Classification of Human- and AI-Generated Texts for English, French, German, and Spanish

Kristina Schaaff and **Tim Schlippe** and **Lorenz Mindner**
IU International University of Applied Sciences, Germany.
kristina.schaaff@iu.org; tim.schlippe@iu.org

Abstract

In this paper we analyze features to classify *human-* and *AI-generated* text for English, French, German and Spanish and compare them across languages. We investigate two scenarios: (1) The detection of text generated by AI from scratch, and (2) the detection of text rephrased by AI. For training and testing the classifiers in this multilingual setting, we created a new text corpus covering 10 topics for each language. For the detection of *AI-generated* text, the combination of all proposed features performs best, indicating that our features are portable to other related languages: The F1-scores are close with 99% for Spanish, 98% for English, 97% for German and 95% for French. For the detection of *AI-rephrased* text, the systems with all features outperform systems with other features in many cases, but using only document features performs best for German (72%) and Spanish (86%) and only text vector features leads to best results for English (78%).

1 Introduction

In recent years, chatbots have gained popularity and are now widely used in everyday life (Pelau et al., 2021). These systems are designed to simulate *human*-like conversations and provide assistance, information, and emotional support (Dibitonto et al., 2018; Arteaga et al., 2019; Falala-Séchet et al., 2019; Adiwardana et al., 2020). OpenAI's ChatGPT has emerged as one of the most commonly used tool for text generation (Taecharungroj, 2023). Within a short span of only five days after its release, over one million users registered (Taecharungroj, 2023). The application scenarios are manifold, ranging from children seeking help with their homework to individuals seeking medical advice or companionship.

As the use of chatbots like ChatGPT becomes more prevalent in our daily lives, it is important to differentiate between *human-generated* and *AI-generated* text. As AI algorithms improve, detecting *AI-generated* content accurately becomes increasingly challenging, posing issues such as plagiarism, fake news generation, and spamming. Thus, tools that can differentiate between *human-* and *AI-generated* content are crucial.

In Mindner et al. (2023), we explored a large number of innovative features such as text objectivity, list lookup features, and error-based features for the detection of English (*EN*) text generated by ChatGPT. However, in the current study, we extended this research to Spanish (*ES*), German (*DE*), and French (*FR*). We selected these languages, as these are amongst the most frequently used languages in the world (Ethnologue, 2023).

Consequently, our contributions are as follows:

- We proved, that the features we investigated in Mindner et al. (2023) can be successfully ported to other languages.
- We extended our *Human-AI-Generated Text Corpus*[1] with *FR*, *DE* and *ES* articles which cover 10 topics, providing a benchmark corpus for the detection of *AI-generated texts* in *EN*, *FR*, *DE* and *ES*.
- Our best systems significantly outperform the state-of-the-art system for the detection of *AI-generated* text ZeroGPT.

2 Related Work

In the this section, we will describe the related work concerning ChatGPT and the classification of *human-* and *AI-generated* texts.

2.1 ChatGPT

Since its release by OpenAI in late 2022, ChatGPT has revolutionized the field of AI (Mesko, 2023) and several other generative AIs such as Google's Bard[2] or Llama[3] (Touvron et al., 2023) have been

[1] https://github.com/LorenzM97/human-AI-generatedTextCorpus
[2] https://bard.google.com
[3] https://ai.meta.com/llama

released. Those tools are capable of generating text in response to user queries across a wide range of domains. Its successful implementation has been demonstrated in areas like education (Baidoo-Anu and Owusu Ansah, 2023), medicine (Jeblick et al., 2022), and language translation (Jiao et al., 2023). ChatGPT is built on the Generative Pre-trained Transformers (GPT) language model and undergoes fine-tuning using reinforcement learning with human feedback. This approach allows ChatGPT to grasp the meaning and intention behind user prompts, enabling it to provide relevant and helpful responses. During the training process, a substantial amount of text data is incorporated to ensure the safety and accuracy of the generated text. While the quantity of training data has not been published, we know that the previous GPT-3 model, which is substantially larger than other language models such as BERT (Kenton and Toutanova, 2019), RoBERTa (Liu et al., 2019), and T5 (Roberts et al., 2019), was trained with 175 billion parameters and 499 billion crawled text tokens (Brown et al., 2020). Through extensive training on a diverse dataset, ChatGPT has acquired a sophisticated understanding of human language, allowing it to generate text that closely resembles that written by humans (Mitrović et al., 2023).

2.2 Detecting *Human-* and *AI-Generated* Texts

Commercial tools and plagiarism apps, such as GPTZero (Shrivastava, 2023), ZeroGPT[4], AI Content Detector[5], and GPT-2 Output Detector[6] (Mitchell et al., 2023), have been developed to identify *AI-generated* text. Furthermore, researchers are working on developing new corpora for this task and finding out which features and classifiers improve classification accuracy: For example, (Yu et al., 2023) present a corpus of *human-* and *AI-generated* abstracts to investigate commercial and non-commercial systems—but only for *EN*. Recent studies have explored various approaches to detect *AI-generated* text, including XGBoost (Shijaku and Canhasi, 2023), decision trees (Zaitsu and Jin, 2023), and transformer-based models (Mitrović et al., 2023; Guo et al., 2023): Mitrović et al. (2023) evaluated characteristics of *AI-generated* text from *EN* customer reviews and built a transformer-based classifier that achieved 79%. Zaitsu and Jin (2023) achieved 100% accuracy in the detection of Japanese texts with decision trees combining stylometric features for Japanese such as bigrams, comma position, and function word rates. Guo et al. (2023) evaluated the characteristics of *human-generated* and *AI-generated* answers to questions in *EN* and Chinese. They fine-tuned a RoBERTa model on their texts and achieved 98.8% F1-score on the *EN* answers and 96.4% F1-score on the Chinese answers. Shijaku and Canhasi (2023) addressed the detection of generated essays written in *EN* and proposed an XGBoost model that achieved 98% accuracy using features generated by TF-IDF and a set of handcrafted features. Soni and Wade (2023) analyzed *human-* and *AI-generated* text summarization and achieved 90% accuracy using DistilBERT[7] (Sanh et al., 2019). Mindner et al. (2023) explored features to detect *AI-generated* and *-rephrased* text for *EN*. They report an F1-score of 96% for *AI-generated* text and 78% for *AI-rephrased* text on their text corpus which contains different topics. These F1-scores were even achieved when the AI was instructed to create the text in a way that a human would not recognize that it was generated by an AI.

To the best of our knowledge, we are the first to explore a large set of features and state-of-the-art classifiers across multiple languages with XGBoost, Random Forrest and MLP. We compare our results with two popular state-of-the-art tools that detect texts generated by AI: GPTZero and ZeroGPT. GPTZero is used by over 1 million people (Shrivastava, 2023), but its results are only reliable for *EN* texts. Consequently, we also used ZeroGPT for comparison which is able to deal with other languages. As there is currently no text corpus available, which contains *human-* and *AI-generated* texts in multiple languages, we extended our *Human-AI-Generated Text Corpus* to cover *EN*, *FR*, *DE* and *ES*.

3 Our Human-AI-Generated Text Corpus

As mentioned in the previous section, we extended our *Human-AI-Generated Text Corpus* (Mindner et al., 2023) to cover *EN*, *FR*, *DE*, and *ES*. In total, for each language we used 100 *human-generated*, 100 *AI-generated*, and 100 *AI-rephrased* articles for our multilingual analysis which contain the following 10 topics: *biology*, *chemistry*, *geography*, *history*, *IT*, *music*, *politics*, *religion*, *sports*, and *visualarts*.

[4]https://www.zerogpt.com
[5]https://copyleaks.com/ai-content-detector
[6]https://openai-openai-detector–mqlck.hf.space
[7]https://huggingface.co/docs/transformers/model_doc/distilbert

Language	Human			AI-generated			AI-rephrased		
	P	S	W	P	S	W	P	S	W
EN	415	1.7k	38.3k	555	1.4k	27.6k	255	1.1k	24.6k
FR	415	1.2k	31.0k	524	1.3k	26.5k	157	0.8k	18.7k
DE	335	1.2k	20.5k	529	1.4k	22.9k	256	1.0k	16.4k
ES	450	1.4k	38.0k	514	1.2k	26.8k	190	0.8k	18.9k

Table 1: *AI-Generated/Rephrased* Text
(P = #paragraphs, S = #sentences, W = #words).

The characteristics of our *Human-AI-Generated Text Corpus* for the respective languages are summarized in Table 1: *EN* consistently has the highest counts across all categories and types of text. On the other hand, the counts for *FR*, *DE*, and *ES* vary substantially depending on whether the text was *human-generated*, *AI-generated*, or *AI-rephrased*. This illustrates how languages differ in the expression of information. The prompts which we used to receive the *AI-generated* and *AI-rephrased* texts are listed in Table 2.

Lang.	Prompt
Text Generation	
EN	Generate a text on the following topic: <topic>
FR	Rédigez un texte sur le thème suivant: <topic>
DE	Erstelle einen Text zum folgenden Thema: <topic>
ES	Genera un texto sobre el siguiente tema: <topic>
Text Rephrasing	
EN	Rephrase the following text: <topic>
FR	Reformulez le texte suivant: <topic>
DE	Formuliere den folgenden Text um: <topic>
ES	Reformule el siguiente texto: <topic>

Table 2: Prompts used for Generation and Rephrasing

4 Our Features for the Classification of *Human-* and *AI-Generated* Texts

As shown in Table 3, we analyzed 37 features for their suitability to discriminate between *human-* and *AI-generated* text. More details of the features are given in Mindner et al. (2023).

4.1 Perplexity-Based Features

Perplexity is a measure of how well a language model is able to predict a sequence of words. The lower the perplexity, the better a language model will perform to predict the next word in a sequence. As *AI-generated* texts are usually based on statistical patterns and rules, they tend to be more repetitive and therefore have a lower perplexity than human generated texts. The *perplexity-based* features in our study are based on the findings by Mindner et al. (2023); Gehrmann et al. (2019); Mitrović et al. (2023); Guo et al. (2023).

For sentence tokenization, we use the Natural Language Toolkit (NLTK)[8]. Perplexity is calculated using *evaluate package*[9] and GPT-2 using the respective models for *EN*[10], *FR*[11], *DE*[12], and *ES*[13].

4.2 Semantic Features

In our study, *semantic* features refer to the properties of words or phrases used to represent their meanings. Previous studies successfully used these features for the differentiation between *human-* and *AI-generated* texts (Mitrović et al., 2023; Guo et al., 2023; Mindner et al., 2023).

Again, we use different Python packages for the respective languages: TextBlob's sentiment analysis for English[14], `textblob-fr`[15] for French, and `textblob-de`[16] for German. Due to the absence of a package that computes both, polarity and subjectivity, for *ES* texts were translated these texts into *EN* using Googletrans[17], despite potential information loss, because of its high BLEU score and proficiency in *ES-EN* translation.

4.3 List Lookup Features

With our *ListLookup* features, we analyze information about the word or character class, e.g., whether it is a stop word or a special character. These features have previously been used for this task by Mindner et al. (2023); Shijaku and Canhasi (2023); Kumarage et al. (2023). For every language, we used ChatGPT to generate a list of all discourse markers as well as the personal pronouns. These lists were additionally evaluated by language experts. To count stop words, we use NLTK for the respective languages.

[8] https://www.nltk.org
[9] https://github.com/huggingface/evaluate
[10] https://huggingface.co/gpt2
[11] https://huggingface.co/dbddv01/gpt2-french-small
[12] https://huggingface.co/dbmdz/german-gpt2
[13] https://huggingface.co/DeepESP/gpt2-spanish
[14] https://textblob.readthedocs.io/en/dev/quickstart.html
[15] https://github.com/sloria/textblob-fr
[16] https://textblob-de.readthedocs.io/en/latest/api_reference.html#module-textblob_de.sentiments
[17] https://github.com/ssut/py-googletrans

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

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

4.4 Document Features

Our *document* features are related to the content and structure of a document such as word frequencies, syntactic structures, and corpus statistics. These features have been successfully used by (Kumarage et al., 2023; Shijaku and Canhasi, 2023; Guo et al., 2023; Mitrović et al., 2023; Zaitsu and Jin, 2023; Mindner et al., 2023). To calculate *sentence-* and *word-related* features, the text is first divided into sentences and words using NLTK's `sent_tokenize` and `word_tokenize` functions. For the features related to Part-of-speech (POS) in *EN* texts, we use the NLTK function `pos_tag`. As NLTK lacks POS tags for the other three languages, we use spaCy NLP library[18]. For POS tags in *DE* texts, we use `de_core_news_sm`[19], for *FR* texts, we use `fr_core_news_sm`[20], and for *ES* texts, `es_core_news_sm`[21].

4.5 Error Based Features

This feature category introduced in Mindner et al. (2023) is based on errors in the text such as grammar and spelling mistakes.

To count multiple blanks, we used regular expressions. Grammar and spelling errors are detected using the open-source tool *LanguageTool*[22] which allows it to detect grammar errors in multiple languages. For the detection of *DE* errors, the built-in class `LanguageToolPublicAPI(de-DE)` for querying the tool's public servers is used. For the other languages, the tool's remote server is applied using the function `Language-Tool(language)`.

[18] https://github.com/explosion/spaCy

[19] https://spacy.io/models/de#de_core_news_sm

[20] https://spacy.io/models/fr#fr_core_news_sm

[21] https://spacy.io/models/es#es_core_news_sm

[22] https://github.com/jxmorris12/language_tool_python

4.6 Readability Features

Readability features assess the readability level of texts as in Mindner et al. (2023); Shijaku and Canhasi (2023); Flesch (1948); Kincaid et al. (1975).

To derive Flesch Reading Ease and Flesch-Kincaid Grade Level we use *Textstat*[23]. This Python library provides functions to calculate text statistics such as grade level, complexity, and readability. Textstat supports calculating Flesch Reading Ease, and Flesch-Kincaid Grade Level for *EN*, *FR*, *DE*, and *ES* texts. However, it is important to note that these measures were originally developed for the specific structure of words, sentences, and syllables of *EN*. Therefore, when applying these measures to texts in *FR*, *DE*, and *ES*, the results may not be as representative as those for *EN*.

4.7 AI Feedback Features

Our *AI Feeback* features reflect, how an AI categorizes the text (Mindner et al., 2023). For this purpose, we use ChatGPT with the prompts in Table 4.

Lang.	Prompt
EN	Was the following text generated by ChatGPT?
FR	Le texte suivant a-t-il été généré par ChatGPT?
DE	Wurde der folgende Text von ChatGPT generiert?
ES	¿El siguiente texto fue generado por ChatGPT?

Table 4: Prompts used for AI Feedback.

4.8 Text Vector Features

Our *TextVector* features analyze semantic content of a text, identifying patterns and repetition (Mindner et al., 2023; Shijaku and Canhasi, 2023; Solaiman et al., 2019; Reimers and Gurevych, 2019).

For the features based on Sentence-BERT, we use the sentence-transformer model `distiluse-base-multilingual-cased-v2`[24], since it supports all the languages used in this research. In addition to the four languages in our experiments, it can be used for more than 50 languages, guaranteeing reliable results for possible future research.

4.9 Summary of Our Analyzed Features

Our 8 feature categories contain 37 features. While the *AI feedback* category consists of one feature, the perplexity, semantic, error-based, and readability features each contain two features. The largest feature category are document features, which contains 19 different features. Table 3 summarizes all the features that are part of our experiments.

[23] https://github.com/textstat/textstat

[24] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2

5 Experimental Setup

In this section, we will describe our experiments with the different feature categories and three classification approaches: The two more traditional approaches XGBoost (Shijaku and Canhasi, 2023) and random forest (RF) (Breiman, 2001) as well as a neural network-based approach with multilayer perceptrons (MLP) (Murtagh, 1991). As in other studies like Guo et al. (2023); Kumarage et al. (2023); Mitrović et al. (2023), we evaluated the classification performance with accuracy (*Acc*) and F1-score (*F1*). First, we built *text generation detection systems* which were trained, fine-tuned, and tested with our *human-generated* and *AI-generated* texts. Second, we implemented *text rephrasing detection systems* which were trained, fine-tuned, and tested with our *human-generated* and *AI-rephrased* texts. To provide stable results, we used a 5-fold cross-validation, randomly dividing our corpus into 80% training, 10% validation, and 10% unseen test set. The numbers in all tables are the average of the test set results. The best performances are highlighted in bold. As a baseline, we choose two popular state-of-the-art tools which detect texts generated by AI: GPTZero and ZeroGPT. GPTZero is used by over 1 million people (Shrivastava, 2023). However, we found that GPTZero's results were only reliable for *EN* texts. Consequently, we used ZeroGPT as our baseline for *FR*, *DE* and *ES*.

6 Results

Table 5 lists *Acc* and *F1* for detecting *AI-generated* and *-rephrased* texts in *EN*, *FR*, *DE*, and *ES*. For each language classifiers trained on *AI-generated* texts achieve better performances compared to classifiers trained on *AI-rephrased* texts.

6.1 Results of Single Feature Categories

As shown in Figure 1 using the example of $sentiment_{subjectivity}$, the distribution of feature values can differ depending on whether the text is *human-generated*, *AI-generated* or *AI-rephrased* and depending on the language. $sentiment_{subjectivity}$ denotes objectivity (low values) or subjectivity (high values) of a text. Average $sentiment_{subjectivity}$ values tend to be higher for *AI-generated* text than for *human-generated* and *AI-rephrased* text. In general, *DE* texts are the most objective texts—be it *human-* or *AI-generated*—while *EN* and *ES* are more subjective. Moreover, *AI-generated* texts tend to be more subjective than *AI-rephrased* texts for our languages.

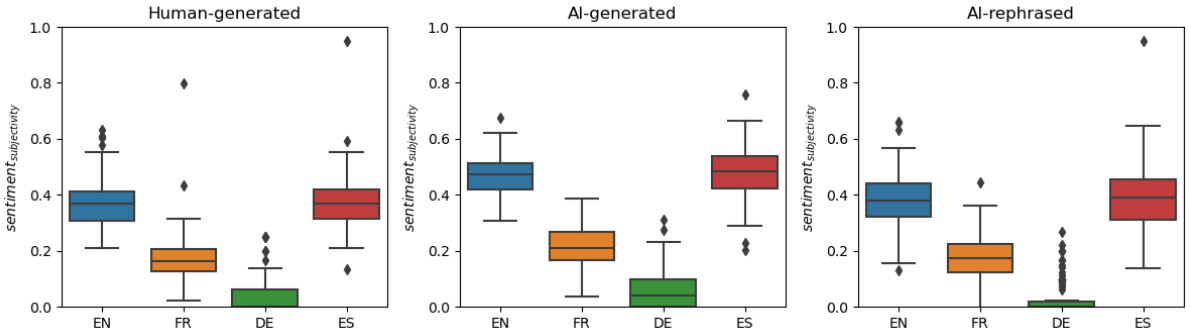

Figure 1: Distribution of $sentiment_{subjectivity}$

		Generated						Rephrased					
		XGBoost		RF		MLP		XGBoost		RF		MLP	
Category	**Lang**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
Perplexity	EN	83.0	82.2	87.0	85.3	82.0	82.1	52.0	48.7	55.0	54.6	56.0	63.2
	FR	62.0	60.3	69.0	66.8	68.0	69.0	50.0	50.2	53.0	44.2	56.0	58.8
	DE	74.0	74.0	76.0	76.1	81.0	80.6	53.0	53.6	61.0	60.4	56.0	62.7
	ES	82.0	82.3	83.0	82.4	82.0	83.6	56.0	55.4	63.0	63.7	62.0	67.3
Semantic	EN	72.0	72.9	75.0	75.6	73.0	72.3	66.0	64.4	66.0	64.3	52.0	54.3
	FR	61.0	55.8	67.0	65.6	63.0	59.4	55.0	48.2	57.0	50.0	51.0	52.9
	DE	64.0	58.3	64.0	59.8	63.0	63.3	56.0	59.9	54.0	54.4	62.0	60.1
	ES	72.0	69.9	75.0	73.8	76.0	75.7	58.0	56.1	58.0	52.4	53.0	56.3
ListLookup	EN	72.0	72.1	79.0	78.5	71.0	67.8	72.0	73.9	67.0	67.5	69.0	70.3
	FR	72.0	73.0	76.0	76.7	67.0	62.9	66.0	62.6	65.0	65.5	64.0	63.2
	DE	74.0	75.8	79.0	77.8	72.0	74.1	57.0	59.1	58.0	59.2	50.0	52.0
	ES	78.0	79.6	82.0	84.1	73.0	76.8	75.0	75.2	80.0	81.3	77.0	78.4
Document	EN	91.0	91.6	92.0	92.6	87.0	86.0	70.0	69.6	71.0	70.8	78.0	76.1
	FR	94.0	94.2	91.0	90.8	92.0	92.2	86.0	85.3	84.0	80.8	81.0	81.2
	DE	87.0	87.2	90.0	89.6	88.0	88.0	**72.0**	**71.9**	67.0	66.7	71.0	71.3
	ES	96.0	96.2	98.0	98.1	87.0	88.5	84.0	83.4	83.0	82.0	**86.0**	**86.4**
ErrorBased	EN	55.0	61.7	55.0	61.7	56.0	63.9	62.0	68.0	62.0	68.0	62.0	68.0
	FR	62.0	64.2	63.0	67.2	61.0	65.5	53.0	56.0	56.0	58.9	56.0	59.7
	DE	67.0	67.1	67.0	67.1	67.0	69.8	62.0	61.9	62.0	63.5	56.0	50.7
	ES	70.0	71.2	71.0	71.9	71.0	74.6	59.0	56.8	61.0	56.3	64.0	65.2
Readability	EN	60.0	56.3	63.0	59.3	60.0	56.8	54.0	51.1	54.0	47.8	50.0	50.2
	FR	61.0	64.7	62.0	66.0	65.0	67.4	59.0	58.3	60.0	60.6	52.0	31.6
	DE	57.0	53.5	53.0	51.5	57.0	53.6	48.0	41.9	45.0	39.1	45.0	44.9
	ES	74.0	73.7	74.0	72.1	69.0	66.6	54.0	49.1	61.0	50.7	56.0	52.5
AIFeedback	EN	62.0	67.1	62.0	67.1	62.0	68.1	52.0	50.9	50.0	39.8	45.0	30.1
	FR	52.0	24.2	52.0	24.2	48.0	37.2	42.0	33.6	42.0	33.6	55.0	53.4
	DE	49.0	46.1	47.0	35.0	50.0	43.4	52.0	61.8	52.0	61.8	50.0	54.3
	ES	52.0	7.3	52.0	7.3	52.0	20.6	50.0	0.0	52.0	7.3	49.0	25.7
TextVector	EN	90.0	89.9	95.0	94.9	83.0	81.7	**79.0**	**78.2**	75.0	71.0	69.0	65.1
	FR	94.0	94.1	93.0	93.0	85.0	85.4	77.0	77.3	75.0	75.2	68.0	64.2
	DE	87.0	87.0	94.0	94.0	90.0	90.8	68.0	67.5	72.0	67.3	72.0	71.7
	ES	84.0	84.5	91.0	89.5	81.0	76.6	76.0	74.0	76.0	73.6	68.0	64.4
All	EN	90.0	90.9	**98.0**	**98.0**	87.0	87.8	77.0	77.6	71.0	69.8	72.0	71.9
	FR	94.0	94.4	**95.0**	**95.0**	88.0	89.2	**89.0**	**87.9**	86.0	84.2	74.0	66.4
	DE	94.0	93.8	**97.0**	**96.9**	87.0	86.6	70.0	71.6	71.0	68.3	70.0	71.6
	ES	94.0	94.4	**99.0**	**99.0**	90.0	90.2	83.0	82.2	83.0	82.9	78.0	76.1

Table 5: Results for the Detection of *EN FR*, *DE* and *ES AI-generated* and AI-Rephrased Texts.

6.1.1 English

Text Generation Detection The results for *EN* in Table 5 indicate that the system that combines all features (*All*) in an RF performs best (*Acc*=98.0%, *F1*=98.0%). The 2nd-best system is the MLP system that uses *Document* features (*Acc*=95.0%, *F1*=94.9%). The RF system that uses *TextVector* features results in a similar per-

formance (*Acc*=95.0%, *F1*=94.9%). The worst-performing system is the XGBoost system that uses the *ErrorBased* features (*Acc*=55.0%, *F1*=61.7%). Compared to GPTZero ($Acc_{GPTZero}$=76.0%, $F1_{GPTZero}$=78.9%), most of our systems perform better. Our best system with all features (*All*) outperforms GPTZero by 28.9% relative in *Acc* and 24.2% relative in *F1*. ZeroGPT reaches 78.0% $Acc_{ZeroGPT}$ and 81.8% $F1_{ZeroGPT}$. Thus, our best system performs 25.6% relatively better in *Acc*, and 19.8% relatively better in *F1*.

Text Rephrasing Detection The performances for the *EN text rephrasing detection systems* are worse than the *text generation detection systems* for all feature categories except *ErrorBased* (*Acc*=62.0%, *F1*=68.0%). The best-performing system is the XGBoost system that uses *TextVector* features (*Acc*=79.0%, *F1*=78.2%), followed by the MLP system that uses *Document* features (*Acc*=78.0%, *F1*=76.1%). The worst-performing system is the MLP system that uses the *AIFeedback* feature. All our *text rephrasing detection systems* were able to outperform GPTZero ($Acc_{GPTZero}$=43.0% and $F1_{GPTZero}$=27.8%). Our the best-performing *TextVector* feature system even outperforms GPTZero by 83.7% relative in *Acc* and even 159.8% relative in *F1*. ZeroGPT reaches 49.0% $Acc_{ZeroGPT}$ and 43.9% $F1_{ZeroGPT}$. Thus, *Document* outperforms it by 61.2% relative in *Acc* and 81.5% relative in *F1*.

6.1.2 French

Text Generation Detection The results for *FR* in Table 5 demonstrate that the system that combines all features (*All*) in an RF performs best (*Acc*=95.0%, *F1*=95.0%). The 2nd-best system is the XGBoost system that uses *Document* features (*Acc*=86.0%, *F1*=85.3%), followed by the XGBoost system that uses *TextVector* features (*Acc*=77.0%, *F1*=77.3%). The worst-performing systems are those that use the *AIFeedback* feature. Our best *FR* system with all features (*All*) outperforms ZeroGPT ($Acc_{ZeroGPT}$=62.0, $F1_{ZeroGPT}$)=72.6%) by 53.2% relative in *Acc* and 30.9% relative in *F1*.

Text Rephrasing Detection The performances for the *FR text rephrasing detection systems* are worse than the *text generation detection systems* for all feature categories except the MLP system that uses the *AIFeedback* feature (*Acc*=55.0%, *F1*=53.4%). The best-performing system is the system that that combines all features (*All*) in an XGBoost (*Acc*=89.0%, *F1*=87.9%), followed by the XGBoost system that uses *Document* features (*Acc*=86.0%, *F1*=85.3%) and the XGBoost system that uses *TextVector* features (*Acc*=77.0%, *F1*=77.3%). The worst-performing systems are again those that use the *AIFeedback* feature. Our best *FR* system with all features (*All*) outperforms ZeroGPT ($Acc_{ZeroGPT}$=57.0, $F1_{ZeroGPT}$)=67.4%) by 56.1% relative in *Acc* and 30.4% relative in *F1*.

6.1.3 German

Text Generation Detection The results for *DE* in Table 5 indicate that the system that combines all features (*All*) in an RF performs best (*Acc*=97.0%, *F1*=96.9%). The 2nd-best system is the RF system that uses *TextVector* features (*Acc*=94.0%, *F1*=94.0%), followed by the RF system that uses *Document* features (*Acc*=90.0%, *F1*=89.6%). As for the previous languages, the worst-performing systems are those that use the *AIFeedback* feature. Our best *FR* system with all features (*All*) outperforms ZeroGPT ($Acc_{ZeroGPT}$=65.0, $F1_{ZeroGPT}$)=70.9%) by 49.2% relative in *Acc* and 36.7% relative in *F1*.

Text Rephrasing Detection The performances for the *DE text rephrasing detection systems* are worse than the *text generation detection systems* for all feature categories except the systems that use the *AIFeedback* features. The best-performing system is the XGBoost system that that uses the *Document* features (*Acc*=72.0%, *F1*=71.9%), followed by the MLP system that uses *TextVector* features (*Acc*=72.0%, *F1*=71.7%). The worst-performing systems are those that use the *Readability* feature. Our best *DE* system with the *Document* features outperforms ZeroGPT ($Acc_{ZeroGPT}$=48.0, $F1_{ZeroGPT}$=49.5%) by 45.5% relative in *Acc* and 45.3% relative in *F1*.

6.1.4 Spanish

Text Generation Detection The results for *ES* in Table 5 show that the system that combines all features (*All*) in an RF performs best (*Acc*=99.0%, *F1*=99.0%). The 2nd-best system is the RF system that uses *Document* features (*Acc*=98.0%, *F1*=89.1%), followed by the RF system that uses *TextVector* features (*Acc*=91.0%, *F1*=89.5%) and the RF system that uses *ListLookup* features (*Acc*=82.0%, *F1*=84.1%). As for the previous languages, the worst-performing systems are those that use the *AIFeedback* feature. The *F1* of 7.3% is so poor since the feature classifies the text as

AI-generated text in almost all cases. Our best *ES* system with all features (*All*) outperforms ZeroGPT ($Acc_{ZeroGPT}$=60.0, $F1_{ZeroGPT}$=71.5%) by 65.0% relative in *Acc* and 38.5% relative in *F1*.

Text Rephrasing Detection The performances for the *ES text rephrasing detection systems* are worse than the *text generation detection systems* for all feature categories. The best-performing system is the RF system that uses the *Document* features (*Acc*=86.0%, *F1*=86.4%). The 2nd best system is the system that combines all features (*All*) in an RF (*Acc*=83.0%, *F1*=82.9%), followed by the RF system that uses the *ListLookup* features (*Acc*=80.0%, *F1*=81.3%). The worst-performing systems are those that use the *AIFeedback* feature. The *F1* of 0% and 7.3% are so poor since the feature classifies the text as AI generated text in almost all cases. Our best *ES* system with the *Document* features outperforms ZeroGPT ($Acc_{ZeroGPT}$=52.0, $F1_{ZeroGPT}$=63.7%) by 65.4% relative in *Acc* and 25.6% relative in *F1*.

6.1.5 Combination of All Features

As shown in Table 5, the best performances for the text generation detection systems are achieved using a combination of all features (*All*). Looking at the systems which use all features, the *Acc* for the *AI-generated FR* and *DE* texts is similar with 97.0%, while the *Acc* for the *AI-generated EN* texts is 98.0%. The best *F1* for the *AI-generated DE* classifier is 96.9%. Thus, it is slightly worse than the classifiers trained on our *EN* and *FR* texts which achieved 98.0% and 97.1%, respectively. The best classifier trained on the *AI-generated ES* texts achieved slightly better performances, with 99.0% *Acc* and 99.0% *F1*. Comparing the performances of the systems trained on the *AI-generated* texts, it can be summarized that the classifiers deliver comparable performances across the languages.

The performances of the systems which use all features (*All*) vary more for the *AI-rephrased* texts across the languages. While the best *EN* classifier reaches 79.0% *Acc* on the *AI-rephrased* texts, the best *FR* classifier achieves 89.0% *Acc* on the *AI-rephrased* texts. The *AI-rephrased* detection system for *DE* only achieves 72.0% *Acc*. Compared to the best *DE* text rephrasing detection system, the *FR* system is 23.6% relatively better in *Acc*. The *Acc* for the *ES* text rephrasing detection system is 1% worse than the *FR* system. For *F1*, comparable conclusions can be drawn across the languages. Thus, our investigated features do not deliver comparable performances for the detection of *AI-rephrased* texts across the evaluated languages.

7 Conclusion and Future Work

In this paper, we investigated features to classify whether text is written by a human, generated by AI from scratch or rephrased by AI. We conducted a comparative analysis of the classification across the languages of *EN*, *FR*, *DE*, and *ES*, assessing the performance of these features in their respective linguistic contexts. To train and test classifiers which use the features, we extended the Human-AI-Generated Text Corpus (Mindner et al., 2023)—our new text corpus, which covers 10 different topics for each of the four languages. For *AI-generated text*, our classifier performed best when combining all features, meaning that there are no substantial differences for features across languages. Therefore, we conclude, that the same feature set could also be used for other languages from the same language families. The accuracies are close with 99% for *ES*, 98% for *EN*, 97% for *DE* and 95% for *FR*. In contrast to that, for the detection of AI-rephrased text, the systems with all features outperformed systems with other features in many cases. For *DE* (72%) and *ES* (86%) we achieved the best results using only document features while for *EN* the text vector features yielded the best results (79%).

Although our results indicate that the same feature set could be applied to other languages within the same familie, future work could investigate the applicability of these features across further language families. This would help in understanding the robustness of our method across a more diverse set of languages. Moreover, our corpus currently covers 10 different topics for each language. Extending the corpus to include more topics, and possibly considering different domains and genres, may help in generalizing the findings and making the system more robust. Finally, experimenting with different machine learning architectures such as transformer models could potentially lead to further optimizations.

Ethics Statement

The collected corpus is made freely available to the community. It is based on Wikipedia and news texts. The research was conducted transparently, free from bias and in compliance with applicable laws and regulations. The use of AI models and data is intended to foster a deeper understanding of AI-generated content, with the goal of promoting responsible use and technological innovation.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V Le. 2020. Towards a Human-Like Open-Domain Chatbot. *ArXiv Preprint ArXiv:2001.09977*.

David Arteaga, Juan Arenas, Freddy Paz, Manuel Tupia, and Mariuxi Bruzza. 2019. 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)*, pages 1–6. IEEE.

David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *Available at SSRN 4337484*.

Leo Breiman. 2001. Random Forests. *Machine Learning*, 45(1):5–32.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *CoRR*, abs/2005.14165.

Massimiliano Dibitonto, Katarzyna Leszczynska, Federica Tazzi, and Carlo M Medaglia. 2018. 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*, pages 103–116. Springer.

Ethnologue. 2023. What are the top 200 most spoken languages?

Clara Falala-Séchet, Lee Antoine, Igor Thiriez, and Catherine Bungener. 2019. OWLIE: A Chatbot that Provides Emotional Support for Coping With Psychological Difficulties. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 236–237.

Rudolf Franz Flesch. 1948. A New Readability Yardstick. *The Journal of applied psychology*, 32 3:221–233.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical Detection and Visualization of Generated Text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Ricke, and Michael Ingrisch. 2022. ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports. *ArXiv E-Prints*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT a Good Translator? A Preliminary Study. *ArXiv Preprint ArXiv:2301.08745*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.

Tharindu Kumarage, Joshua Garland, Amrita Bhattacharjee, Kirill Trapeznikov, Scott Ruston, and Huan Liu. 2023. Stylometric Detection of AI-Generated Text in Twitter Timelines.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Bertalan Mesko. 2023. The chatgpt (generative artificial intelligence) revolution has made artificial intelligence approachable for medical professionals. *Journal of medical Internet research*, 25:e48392.

Lorenz Mindner, Tim Schlippe, and Kristina Schaaff. 2023. Classification of Human- and AI-Generated Texts: Investigating Features for ChatGPT. *TBD*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*.

Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-Generated Text. *arXiv preprint arXiv:2301.13852*.

Fionn Murtagh. 1991. Multilayer Perceptrons for Classification and Regression. *Neurocomputing*, 2(5):183–197.

Corina Pelau, Dan-Cristian Dabija, and Irina Ene. 2021. 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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, Katherine Lee, Michael Matena, Noam Shazeer, Peter J. Liu, Sharan Narang, Wei Li, and Yanqi Zhou. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Technical report, Google.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. In *Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS Edition (EMC2-NIPS 2019)*.

Rexhep Shijaku and Ercan Canhasi. 2023. ChatGPT Generated Text Detection.

Rashi Shrivastava. 2023. With Seed Funding Secured, AI Detection Tool GPTZero Launches New Browser Plugin.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release Strategies and the Social Impacts of Language Models.

Mayank Soni and Vincent Wade. 2023. Comparing Abstractive Summaries Generated by ChatGPT to Real Summaries Through Blinded Reviewers and Text Classification Algorithms.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models.

Peipeng Yu, Jiahan Chen, Xuan Feng, and Zhihua Xia. 2023. CHEAT: A Large-scale Dataset for Detecting ChatGPT-writtEn AbsTracts.

Wataru Zaitsu and Mingzhe Jin. 2023. Distinguishing ChatGPT(-3.5, -4)-generated and Human-Written Papers Through Japanese Stylometric Analysis.