MAiDE-up: Multilingual Deception Detection of GPT-generated Hotel Reviews

Anonymous ACL submission

Abstract

Deceptive reviews are becoming increasingly common, especially given the increase in performance and the prevalence of LLMs. While work to date has addressed the development of models to differentiate between truthful and deceptive human reviews, much less is known about the distinction between real reviews and AI-authored fake reviews. Moreover, most of the research so far has focused primarily on English, with very little work dedicated to other languages. In this paper, we compile and make publicly available the MAiDE-up dataset, consisting of 10,000 real and 10,000 AI-generated fake hotel reviews, balanced across ten languages. Using this dataset, we conduct extensive linguistic analyses to (1) compare the AI fake hotel reviews to real hotel reviews, and (2) identify the factors that influence the deception detection model performance. We explore the effectiveness of several models for deception detection in hotel reviews across three main dimensions: sentiment, location, and language. We find that these dimensions influence how well we can detect AI-generated fake reviews.

1 Introduction

Recent advancements in Natural Language Generation (NLG) technology have greatly improved the quality of LLM-generated text. One notable example is OpenAI's ChatGPT, which has demonstrated exceptional performance in tasks such as story generation (Yuan et al., 2022), question answering (Bahak et al., 2023), essay writing (Stokel-Walker, 2022), and coding (Becker et al., 2023). However, this newfound ability to produce highly efficient, human-like texts also raises concerns about detecting and preventing the misuse of LLMs (Pagnoni et al., 2022; Mirsky et al., 2023).

One particular problem is the prevalence of AI-generated reviews, and while tools and datasets have been proposed, none have solved the problem completely (Wu et al., 2023b), which in turn, leads to the erosion of trust in online opinions. Furthermore, most of the research so far has focused primarily on English, with very little work dedicated to other languages. Our study aims to fill in these gaps and provide novel insights into multilingual LLM-generated hotel reviews. We analyze different types of human-interpretable features, such as linguistic style, writing structure, topics, and psycholinguistic markers, along with baselines across multiple models. We hope our research will help organizations leverage NLP to combat the use of LLMs in scenarios where genuine, human-generated content is highly valued, such as customer reviews on platforms like Booking, Yelp, and Amazon.

Specifically, our paper aims to answer two main research questions.

RQ1: What are the linguistic markers (syntactic and lexical) of multilingual LLM-generated reviews when compared to human-authored reviews?

RQ2: Which factors influence multilingual deception detection performance?

The paper makes the following contributions. First, **we compile and share a multilingual dataset of 10,000 real and 10,000 AI-generated fake hotel reviews, balanced across ten languages:** Chinese, English, French, German, Italian, Korean, Romanian, Russian, Spanish, and Turkish, as well as across ten locations and different sentiment polarities. To the best of our knowledge, this is the first dataset of multilingual reviews at this scale. Second, using this dataset, **we conduct extensive linguistic style and lexical analyses to compare the AI-generated deceptive hotel reviews with the real human-written hotel reviews.** Finally, **we explore the effectiveness of different models for multilingual deception detection** in hotel reviews across three dimensions: sentiment, location, and language.

2 Related Work

LLM-generated Text Detection. The powerful generation capabilities of LLMs has made it challenging for humans to differentiate between LLM-generated and human-written texts (Jakesch et al., 2023). This led to extensive research being conducted on developing models for detecting LLM-generated text, including fine-tuned LMs (Jawahar et al., 2020; Guo et al., 2023), zero-shot methods (Solaiman et al., 2019; Ippolito et al., 2019), watermarking techniques (Kirchenbauer et al., 2023a,b), adversarial learning methods (Hu et al., 2023; Koike et al., 2024), LLMs as detectors (Bhattacharjee and Liu, 2024; Yu et al., 2023), and human-assisted methods (Dou et al., 2021).

Fine-tuning LMs, in particular Roberta (Liu et al., 2019), has had great success in binary classification settings (Fagni et al., 2021; Radford et al., 2019), and therefore, we also adopt it in our work. On average, these models yielded a 95% accuracy, outperforming zero-shot, and watermarking methods and showing resilience to various attacks within in-domain settings. However, just like other encoder-based fine-tuning approaches, these models lack robustness when dealing with cross-domain or unforeseen data (Bakhtin et al., 2019; Antoun et al., 2023).

Our work falls under the category of black-box modeling, as described in a recent survey of AI language detectors by Tang et al. (2023). We are interested in the outputs of LLMs, rather than the specific details of a model's contents or design. This approach allows us to focus on the differences between human and AI-generated texts, instead of the particular implementation details of models.

Human vs. LLM-generated Text. There have been several efforts to study the differences between AI and human-generated text (Sadasivan et al., 2023; Jakesch et al., 2023; Markowitz et al., 2024; Markowitz and Hancock, 2024)

In the context of deception, Giorgi et al. (2023) and Markowitz and Hancock (2024) argue that AI-generated text is *inherently deceptive* when describing human experiences like writing reviews because the system is not grounded in material world experiences. At the same time, Giorgi et al. (2023) and Markowitz and Hancock (2024) are the closest to our work. They use ChatGPT to generate hotel reviews and compare them to human-written deceptive and truthful hotel reviews from TripAdvisor. This data is collected by Ott et al. (2011) from 20 hotels in Chicago, IL, USA. Markowitz and Hancock (2024) find that AI-generated text has a more analytic style and is more affective, more descriptive, and less readable than human-written text, while Giorgi et al. (2023) find that human-written text is more diverse in its expressions of personality than AI-generated text. We replicate their style analysis findings and extend the data collection and analysis to nine other languages and ten global hotel locations. Unlike prior work, we do not analyze deceptive human reviews. In addition, we also analyze the role of sentiment, language, and location in deception performance and implicitly in the quality of the AI-generated reviews.

Multilingual LLM-generated Text Detection. While several multilingual models (Tulchinskii et al., 2024; Mitchell et al., 2023; Antoun et al., 2023; Guo et al., 2023) and datasets (Wang et al., 2023; Guo et al., 2020) have been proposed, Wu et al. (2023a), in their comprehensive survey on LLM-generated text, address the need for multilingual datasets and models to facilitate the evaluation of text detectors generated by LLMs across different languages. Addressing them is essential for the usability and fairness of detectors for LLM-generated text. In our work, we address this gap, by providing a dataset, analysis, and classification models for 10 languages.

3 The MAIDE-UP Dataset

To answer our research questions, we compile a novel dataset, which we refer to as MAIDE-UP - Multilingual Ai-generateD fakE reviews. MAIDE-UP contains a total of 20,000 hotel reviews: 10,000 are real, human-written, and 10,000 are fake, LLM-generated. The reviews are balanced across language, location, and sentiment. We outline our process for collecting real and fake reviews below.

3.1 Real Hotel Reviews

We collect 10,000 hotel reviews from Booking.com, [1] which is one of the largest marketplaces for online travel bookings. The data is balanced per language, location, and sentiment. Finally, we ensure data quality through automatic and manual assessments.

Languages. The dataset is balanced across ten languages: *Chinese, English, French, German, Italian, Korean, Romanian, Russian, Spanish*, and *Turkish*. We automatically web-crawl Booking.com for each of the ten languages. Additional

[1] booking.com

details on how we automate this process can be found in Appendix A.1.

Locations. We collect reviews from hotels located in popular capital cities: *Ankara, Beijing, Berlin, Bucharest, Madrid, New Delhi, Paris, Rome, Seoul*, and *Washington*. Most of the cities are selected to be the capitals of countries where the official language is one of the 10 languages (New Delhi and Moscow are the exceptions). To ensure the collection of an equal number of reviews for each language, particularly for *Chinese, Korean, Romanian*, and *Turkish*, where the number of reviews per hotel is often limited, we identify 250 hotels in each city and collect up to 50 reviews per hotel.

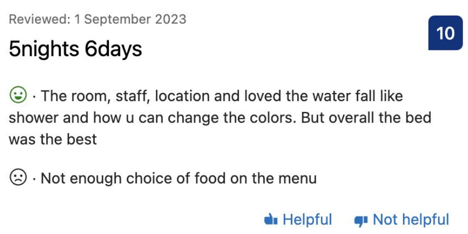

Figure 1: An example of an English positive review, rated with a score of 10, that contains both "upside" and "downside" sections. The reviewer can choose to write in just one or both sections.

Sentiment. The Booking.com platform provides *review scores* from 1 to 10, where a score of 1-6 represents a negative review and 7-10 is a positive review.[2] We collect a balanced number of positive and negative reviews for each language, i.e., 500 positive and 500 negative for each language. The platform provides a specific review format consisting of two parts, "upside" and "downside", to allow users to separate their positive and negative feedback. The "upside" and "downside" sections are optional, as the reviewer can write in just one or both sections. Figure 1, shows a review example of a positive review, rated with a score of 10, that contains both "upside" and "downside" sections.

Quality Assurance. We automatically verify each review's language and filter out reviews in a different language than the one crawled for (more information in Appendix A.1). To further ensure the data quality, ten native speakers manually verified 50 random reviews each. Specifically, they verified the review's syntax and semantics and ensured the sentiment aligned with the content for the "upside" and "downside" parts of the review, as well as for the review overall. One potential concern is the possibility of fake human-written reviews. Booking.com addresses this by combining specialized personnel and automated systems to detect and remove fake reviews.

3.2 LLM-generated Hotel Reviews

We generate 10,000 hotel reviews with GPT-4.[3] The generated reviews follow the same distribution across language, location, and sentiment, as the real hotel reviews, as described in section 3.1. We use GPT-4 because it is one of the largest LLMs available and has been demonstrated to effectively emulate human texts (Achiam et al., 2023).

3.2.1 Prompt Design and Robustness

GPT-4 takes a prompt as input, which is comprised of a list of *message* objects, and returns one generated hotel review as output. We use *messages*, which are more interactive and dynamic compared to the classical prompt-style. Specifically, we use messages with two properties: *role* and *content*: the role takes one of three values: "system", "user" or "assistant", while the content contains the text of the message.[4]

The prompt is first formatted with a "system message" role, which sets the behavior of the model. This is followed by two rounds of conversations between the roles of "assistant" and "user" in a *few-shot prompting* technique. Finally, we use a "user message" to prompt the model to generate a hotel review in a specified language, location (hotel name and capital city), and sentiment.

System Prompt. We find that we can obtain high-quality responses with additional context in our prompts. Therefore, we instruct the model to be a well-traveled native {language} tourist in the "system message". The {language} placeholder is replaced by the language name of the hotel reviews we aim to collect. The instructions also contain information about the language, location (hotel name and city) and the sentiment of the review, as well as the output format, as illustrated below:

> *You are a well-traveled native {language} tourist, working on writing hotel reviews of hotels you have stayed in. Given hotel name, city name, language, and sentiment, you write a hotel review comprised of upside and downside. Then*

[2] Even though it is marked as 'Pleasant' on Booking.com, a score of 6 is considered negative on the platform.

[3] https://platform.openai.com/docs/guides/text-generation/chat-completions-api

[4] https://help.openai.com/en/articles/7042661-chatgpt-api-transition-guide

you give an overall review score, an integer ranging from 1 to 10 where the score larger than 6 indicates positive experience, otherwise negative experience. You always output a JSON containing the following keys: "Upside_Review", "Downside_Review", "Review_Score". Reviews are always in consistent styles, tone, sentence structure.

User Prompt. To increase the diversity and robustness of the generated data, we collect 36 multilingual "user messages". Specifically, we ask ten native speakers to each write four "user messages": two in their own language and two in the corresponding *English* translation. The guidelines given to the annotators are shown in Appendix A.2.

We use the "user messages" to prompt GPT-4 to generate a hotel review with a specified language, hotel location, and sentiment. Below is an example of a "user message" in *Spanish* and its *English* translation. We show all the multilingual "user messages" in the Appendix Table 6.

¿Podés escribir un comentario positivo en L sobre el hotel H de C? Can you please write a positive review in L for the hotel H located in C?

Few-shot Prompting. We use the conversations between "user" and "assistant" when generating data. The "user messages" are randomly selected from our multilingual "user messages" and the "assistant messages" are randomly extracted from our collected real hotel reviews. The answer to the last "user message" is automatically generated by the "assistant message" and used to collect the GPT-4 generated hotel reviews.

Quality Assurance. To ensure the quality of our generated data, we conduct sanity checks by asking native speakers to review approximately 100 hotel reviews in their respective languages. The reviews are checked for readability, syntax errors, and style. Based on the feedback, we find that some of the generated *Chinese* reviews contain nonsensical phrases. Additionally, reviews in other languages such as *Romanian, Korean*, and *Spanish* have a formal tone. Table 1 displays random reviews in *English*, both real and generated, with positive and negative sentiments.

Cost. We generated 10,000 posts in 10 languages for a total cost of 250-300 dollars (0.03 per 1K input tokens and 0.06 per 1K output tokens).

Real Hotel Reviews
+ *Nicely furnished room and nicely decorated lobby. Room service is affordable and the receptionists, especially Mr. Omar, are usually eager to help with any queries; N/A*
- *The location was very central and the staff was nice and helped us.; The room smelled like cigarettes and there was mold in the bathroom*
LLM-generated Hotel Reviews
+ *The staff was extremely helpful and accommodating. Clean and well-furnished rooms. Central location with easy access to public transport.; The breakfast offerings could be more varied.*
- *Nothing. Horrible experience.; Bad customer service, rooms were not clean and the food was below average.*

Table 1: Random review samples with positive (+) and negative (-) sentiments in English. "Upside" and "Downside" parts are separated by ";". Multilingual sampled reviews are shown in Appendix Tables 7 and 8.

4 Multilingual Analyses of Real and LLM-generated Hotel Reviews

Using our dataset, we conduct extensive linguistic analysis to compare the AI-generated fake hotel reviews with the real human-written hotel reviews.

Analytic Writing. This is an index that measures the complexity and sophistication of the writing, which can be an indicator of advanced thinking. This technique has been used in various fields, including persuasion (Markowitz, 2020), analysis of political speeches (Jordan et al., 2019), and gender studies (Meier et al., 2020), among others. The formula for analytic writing is $[articles + prepositions - pronouns - auxiliaryverbs - adverb - conjunctions - negations]$ from LIWC scores (Jordan et al., 2019; Pennebaker et al., 2014). We use Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2007, 2015), a gold-standard text analysis tool, to obtain the categories in the index formula for the following languages: *Chinese*, *English*, *French*, and *Spanish*.[5] Since they are translations for different versions of the LIWC lexicon, e.g., 2001, 2007, and 2015, we first align them using the intersection with the 2015 English version. We find that **AI-generated texts in *English* are more complex than real texts**, which aligns with our observations from data quality checks. The results for *Chinese*, *French*, and *Spanish* are not statistically significant (see Table 2).

[5]While versions for Korean and Turkish are listed as available upon request, we were unable to obtain them.

Descriptiveness. The descriptiveness of a text can be measured by its ratio of adjectives, as texts with high rates of adjectives tend to be more elaborate and narrative-like compared to texts with low rates of adjectives. (Chung and Pennebaker, 2008) Additionally, adjectives are often used in false speech, making them a key marker of deceitful language (Johnson and Raye, 1981). We measure the ratio of adjectives using the multilingual library `textdescriptives` from (Hansen et al., 2023).[6] For *Turkish* we use `HuggingFace` from Altinok (2023).[7] In Table 2, we show that **AI-generated text is usually more descriptive than real text**. The only exceptions are *German* reviews, where the real text is more descriptive, and *Korean* reviews, where the difference is not significant.

Readability. The readability of a text is reflected not only by its word count, but also by the word complexity, with e.g., longer words being more difficult to read and understand. We use the Flesch Reading Ease metric (Flesch, 1948), which counts the number of words per sentence and syllables per word. This metric is used to assess the structural complexity of language patterns in various texts, such as scientific articles (Markowitz and Hancock, 2016), online petitions (Markowitz, 2023), and social media data (Hubner and Bond, 2022). We use the multilingual library `textdescriptives` to compute the Flesch Reading Ease metric and the word count. In Table 2, we find that **AI-generated text is usually less readable and more wordy than real text**. The only exceptions are *German* and *Russian*, with no significant differences.

Topic Modeling. We compute the most prevalent topics and their keywords with a multilingual pipeline from Scattertext (Kessler, 2017). Each review is processed to obtain the most important words with high TF-IDF scores (Ramos et al., 2003). We pre-process the text with `spaCy` multilingual pipelines to tokenize hotel reviews and lowercase tokens, remove stop words, and lemmatize tokens.[8]

Comparing real to generated reviews, we find that in multilingual data, the word "Booking" is most frequent in real hotel reviews, along with words like "reception (ro: receptie)", "checking", "bathroom (ro: baie, es: bagno)", "shower". In contrast, generated hotel reviews contain more words about "service", "comfort (de: komfortabel, ch: 舒适)" and "room (ro: camerele)". In English hotel reviews, AI-generated reviews contain more mentions of city names: "Bucharest", "Washington", "Ankara", while real hotel reviews contain more words about "cleaning" and "time". The topic distribution across real and AI-generated hotel reviews in all languages is shown in Figure 9 and Figure 8 in the Appendix.

5 Multilingual Deception Detection

We explore the effectiveness of different models for multilingual deception detection. As interpretable baselines, we train and test a Naive Bayes (Lewis, 1998) and a Random Forest (Ho, 1995) classifier. As our main model, we fine-tune an XLM-RoBERTa (Conneau et al., 2019) classifier, which lacks interpretability but is highly performant.

Model Training Setup. We use one model for all the multilingual data and also experiment with different training and test data splits. Specifically, we use a *default* and a *few-shot* train-test split as shown in Table 3. Note that the *few-shot* setting closely resembles the real-life scenario when a user who wants to generate fake multilingual reviews has access to little labeled data.

5.1 Interpretable Baselines

We extract multiple interpretable features for each review: token counts, TF-IDF scores (Ramos et al., 2003), as well as the Analytic Writing, Descriptiveness, and Readability scores described in Sec 4.

Random. The reviews are split equally between real and generated; therefore, a random baseline has an accuracy of 50% and an F1 score of 0%.

Naive Bayes. As a simple and interpretable baseline, we train a Naive Bayes classifier (Lewis, 1998) to distinguish between real and AI-generated reviews. We use the Gaussian Naive Bayes model from the `sklearn` (Pedregosa et al., 2011) library with the default settings.

Random Forest. We train a Random Forest classifier (Ho, 1995) to distinguish between real and AI-generated reviews. We use the default model from the `sklearn` (Pedregosa et al., 2011) library.

[6] https://github.com/HLasse/TextDescriptives
[7] https://huggingface.co/turkish-nlp-suite
[8] https://spacy.io/models

Lang	Analytic writing			Descriptiveness			Readability			Word Count		
	Real	Gen	Diff	Real	Gen	Diff	Real	Gen	Diff	Real	Gen	Diff
🇨🇳	6.2 ± 7.0	6.5 ± 3.6	0.3	2.6 ± 3.9*	4.2 ± 3.5*	1.6	-	-	-	62.1 ± 92.9*	79.4 ± 41.2*	17.3
🇺🇸	11.9 ± 6.8*	18.6 ± 5.2*	6.7	13.7 ± 8.6*	15.5 ± 5.5*	1.8	66.3 ± 29.0*	54.6 ± 17.1*	-11.7	49.2 ± 54.7*	55.7 ± 28.5*	6.5
🇫🇷	21.0 ± 10.1	20.8 ± 6.2	-0.2	13.9 ± 9.3*	16.2 ± 6.7*	2.3	51.3 ± 29.9*	43.9 ± 15.8*	-7.4	39.6 ± 43.5*	47.5 ±24.2*	7.9
🇩🇪	-	-	-	7.0 ± 7.4*	5.4 ± 5.5*	-1.6	28.2 ± 32.3	27.2 ± 19.3	-1	41.2 ± 43.0*	47.0 ± 22.4*	5.8
🇮🇹	-	-	-	15.4 ± 10.2*	19.3 ± 7.6*	3.9	-9.4 ± 32.0*	-13.7 ± 17.9*	-4.3	37.9 ± 41.7*	45.0 ± 23.2*	7.1
🇷🇴	-	-	-	11.3 ± 9.1*	15.5 ± 6.3*	4.2	8.4 ± 36.7*	-0.6 ± 19.3*	-9	39.6 ± 46.3*	44.1 ± 22.1*	4.5
🇰🇷	-	-	-	6.0 ± 6.4	5.7 ± 4.3	-0.3	-	-	-	30.1 ± 31.7*	32.6 ± 14.1*	2.5
🇷🇺	-	-	-	13.7 ± 8.7*	18.9 ± 6.0*	5.2	-8.2 ± 35.1*	-16.8 ± 20.0*	-8.6	42.8 ± 45.5	40.7 ± 20.5	-1.1
🇪🇸	12.9 ± 7.0	15.9 ± 4.1	3	11.5 ± 8.6*	14.9 ± 5.4*	3.4	19.9 ± 27.6*	11.0 ± 17.1*	-8.9	39.0 ± 40.5*	48.0 ± 23.1*	9
🇹🇷	-	-	-	13.6 ± 9.5*	14.4 ± 6.0*	0.8	-	-	-	26.4 ± 25.3*	32.8 ± 14.7*	6.4
Average	13± 7.7	15.4±4.7	2.4	10.8 ± 8.1	13 ± 5.6	2.1	22.3 ± 31.8	15 ± 18.1	-7.2	40.7 ± 46	47.2 ± 23	6.5

Table 2: To what degree is AI-generated text different from real text in terms of analytic writing, descriptiveness, readability, and word count? We compute the mean and standard deviation for all the reviews, across each language. We mark (*) when the difference between real and generated data is statistically significant, based on the Student t-test (Student, 1908) with p-value < 0.05. The significant differences are indicated in *teal* when the generated data scores are higher than real data scores, and in *olive* otherwise.

	default		*few-shot*	
	real	gen	real	gen
Train	8,000	8,000	100	100
Test	2,000	2,000	9,900	9,900

Table 3: Number of reviews for the experimental data split. The *default* split corresponds to an 80-20% train-test data split, while the *few-shot* split corresponds to a 1-99% train-test data split.

5.2 Main Model

XLM-RoBERTa. We fine-tune XLM-RoBERTa base (Conneau et al., 2019) model from `HuggingFace`[9]. The model is a multilingual version of RoBERTa (Liu et al., 2019), and is pre-trained on 2.5TB of filtered CommonCrawl (Wenzek et al., 2020) data containing 100 languages.

We use a learning rate of $2e-5$ and a batch size of 8. We train for 5 epochs and take the best epoch based on validation accuracy. The validation data represents 10% of the train data set.

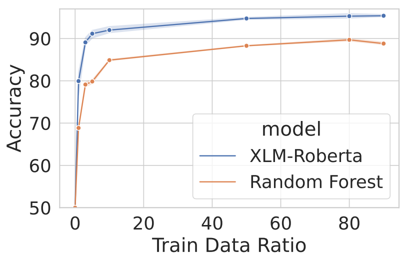

Figure 2: Accuracy measured with XLM-RoBERTa and best Random Forest model on different ratios of training data. The accuracy plateaus at 10%, i.e., 2,000 reviews.

[9] https://huggingface.co/FacebookAI/xlm-roberta-base

Features	Setup	Accuracy	F1
Interpretable Baselines: NAIVE BAYES / RANDOM FOREST			
Analytic scores	*default*	53.3 / 53.9	63.2 / 64.3
	few-shot	52.1 / 51.1	63.6 / 32.1
Descriptive scores	*default*	58.3 / 61.0	63.1 / 58.6
	few-shot	58.4 / 57.4	57.7 / 57.5
Readability scores	*default*	53.5 / 57.3	61.6 / 58.7
	few-shot	53.6 / 53.0	61.8 / 57.7
Token counts	*default*	79.6 / 87.4	82.2 / 88.0
	few-shot	54.4 / 62.9	54.9 / 66.6
TF-IDF scores	*default*	84.5 / 87.7	83.9 / 86.9
	few-shot	64.6 / 67.1	69.6 / 57.1
All scores	*default*	84.3 / **89.3**	84.1 / **89.2**
	few-shot	66.3 / 70.6	70.5 / 66.8
Main Model: XLM-ROBERTA			
-	*default*	**94.8**	**94.9**
-	*few-shot*	76.6	80.1
Human	-	71.5	69.1

Table 4: Classification test results with the *few-shot* and *default* setups over all languages.

6 Evaluation

We show the Accuracy and F1 results for all the models in Table 4. The model with the best detection performance is XLM-RoBERTa, with an accuracy of 94.8% on the *default* train-test data split and 76.6% accuracy on *few-shot* train-test split.

Among all the interpretable models, Random Forest with all the features achieves the best accuracy of 89.3 on the *default* train-test data split and 70.6% accuracy on the *few-shot* train-test data split. **The features that contribute the most to the performance increase are TF-IDF scores, followed**

Most salient words for the Real reviews	Most salient words for the Generated reviews
NUM: 20, 10, 15, 100, 30, 50, tres (three,), dos (two,)	**VB & ADV:** желать (want,), 또한 (also,), lasciava (left,), 않았습니다 (however,), ayudar (help,)
ADJ: yüzlü (faced,), ok (-,), cok (a lot,), min (-,), noi (new,), 좋아요 (great,)	**ADJ:** отеляялучшего (the best,), enttäuschend (disappointing,), 만족스럽지 (not satisfied,), limpias(clean,), отелядружелюбным (friendly,), dispuesto (willing,), 위치해 (located,), prietenos (friendly,), 편리했습니다 (convenient,), negative (negative,), lent (slow,)
NOUN: check (-,), 가성 (cost-effectiveness,), habitacion (room,), ubicacion (location,), güler (laughs,), euro (-,), booking (-,), 그런지 (grunge,), terlik (slipper,), kahvalti (breakfast,), minuti (minutes,), минут (minutes,)	**NOUN:** serviciul (service,), seçenekleri (options,), bağlantısı (connection,), wifiul (the wifi,), oraș (city,), kalitesi (quality,), 호텔의 (hotel,)

Table 5: Top 20 most salient features for real and generated reviews from the best Random Forest model on the *default* train-test split. Each word is accompanied by its English translation, part of speech, and language.

by Token counts, Descriptiveness, Readability, and Analytic scores. Note that missing data for several languages, as seen in Table 2, impacts the performance of Analytic scores and Readability scores. To handle the missing values, we replace them with the mean of the present values. Even though they contribute to performance increase along with the TF-IDF scores, we find that the analytic writing, descriptiveness, and readability scores are not sufficient to accurately distinguish between real and LLM-generated hotel reviews.

Both models have a fairly high *few-shot* performance, indicating that they are capable of learning this task from very few examples, i.e., 200 reviews balanced across 10 languages. Furthermore, Figure 2 shows that **both models learn considerably from just 2,000 reviews** (train data ratio of 10%), and increasing the training data does not lead to high performance gains.

Human Evaluation. We ask ten native speakers to manually classify 200 random reviews (100 real, 100 generated) across all languages as deceptive or not. The accuracy per language is shown in Figure 4. The 71.5% overall accuracy suggests that **humans find it moderately difficult to distinguish between real and generated hotel reviews**. Indeed, the annotators mention that the task is challenging, and they tend to label more complex and formal reviews as AI-generated. In Table 4, human performance is comparable to the *few-shot* learning model performance. However, **all the models trained on more data (*default* split) significantly outperform humans**, indicating that they are helpful for deception detection.

6.1 Interpretability Analysis

Table 5 shows the top 20 most salient features of the best Random Forest classifier in the *default* training setup. We observe that *Korean, Spanish,* and *Turkish* words appear most frequently, suggesting they have the strongest impact on deception detection performance. Additionally, for both real and generated reviews, words related to the hotel topic are most salient: *cost, room, location, booking, breakfast, wifi, service*. Compared to generated reviews, real reviews tend to have more numerals as salient words, while generated reviews have more verbs, adverbs, and adjectives. This finding aligns with our previous discovery that when measuring descriptiveness scores, generated reviews are typically more descriptive than real reviews. Finally, the nouns in real reviews tend to be more diverse than those in generated reviews: *cost-effectiveness, laughs, grunge, slipper*. We encourage further research to delve more deeply into the lexical variances between real and generated multilingual reviews using our dataset.

6.2 Ablations per Language, Location, and Sentiment

We show the XLM-RoBERTa main model *few-shot* performance across sentiment, review language, prompt language, and hotel location in Figure 3. We use 10-fold cross-validation to compute confidence intervals and compute the significance using Student t-test (Student, 1908) and p-value <0.05.

Review Language. As shown in Figure 3 (a), deception performance is lowest for *Korean* and *English* reviews, which implies that **GPT4 is better at generating *English* and *Korean* reviews**. On the other hand, deception performance is highest for *German* and *Romanian* reviews, indicating that **GPT4 is worse at generating *German* and *Romanian* reviews**. One possible explanation is the quantity of training data accessible for each language. Specifically, we find a moderate correlation (Pearson coefficient of 0.53) between the amount of training data and the GPT-4 performance for each language. The training data is estimated from the CommonCrawl dataset (Wenzek et al., 2020).

Prompt Language. From the perspective of a user interested in generating hotel reviews with

GPT4, we measure how the language of the user prompt impacts the quality of the generated reviews. As shown in Figure 3 (b), deception performance is lowest for *Turkish* and *Korean*, indicating that **GPT4 is better at generating multilingual reviews when prompted in *Turkish* and *Korean*.** On the other hand, deception detection performance is highest for *English* and *French* prompts, thus **GPT4 is worse at generating multilingual reviews when prompted in *English* and *French*.** We measure a high negative correlation (Pearson coefficient of -0.63) between the amount of training data and and the performance of GPT-4 for each language. Therefore, understanding the impact of the language prompt on the quality of generated data is a complex issue that requires further investigation in future work.

Hotel Location. When generating multilingual hotel reviews using multilingual prompts, we also measure how the location of the hotel impacts the quality of the generated reviews. As shown in Figure 3 (c), deception performance is lowest for *Seoul*, *Rome*, and *Beijing*, indicating that **GPT4 is better at generating multilingual reviews for hotels in *Seoul*, *Rome*, and *Beijing*.** On the other hand, deception performance is highest for *Bucharest*, *Washington*, *Ankara*, and *Berlin*, thus **GPT4 is worse at generating multilingual reviews for hotels in *Bucharest*, *Washington*, *Ankara*, and *Berlin*.** The results per location are moderately correlated (Pearson coefficient of 0.46) with the results per language review. Therefore, the amount of training data available for each language might also affect the model's ability to generate reviews about the locations where that language is spoken.

Review Sentiment. Across sentiment polarities, our main deception classification model obtains an accuracy of 79.9 for positive reviews and 82.6 for negative reviews. The performance difference between the two types of reviews is statistically significant. A lower deception classification performance on positive reviews indicates that **GPT4 is more proficient in generating multilingual positive reviews than negative reviews**. This is expected given that until recently, the model did not allow negative reviews to be produced, and it also tends to be less negative than human-authored text. (Markowitz et al., 2024)

7 Conclusion

In this paper we focused on the identification of AI-written fake hotel reviews in multiple languages.

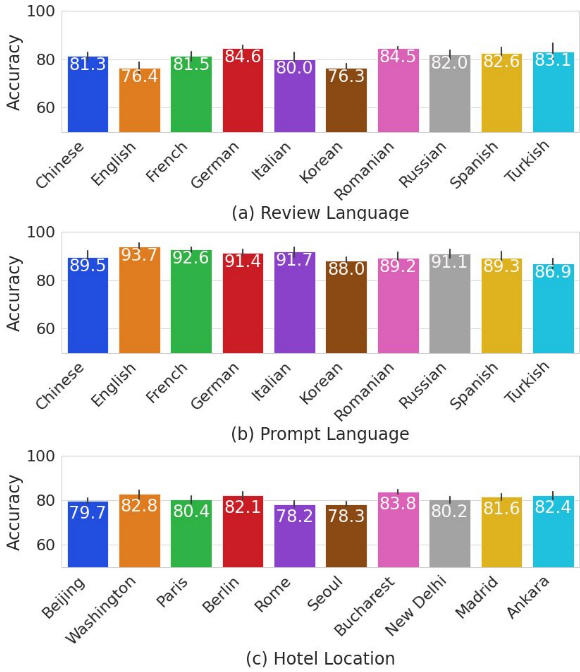

Figure 3: Accuracy with XLM-RoBERTa model per (a) review language, (b) prompt language, and (c) hotel location on the *few-shot* train-test split.

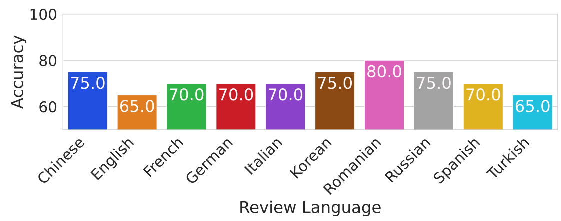

Figure 4: Human accuracy per review language.

To facilitate research in this domain, we released MAIDE-UP, a dataset of 10,000 real and 10,000 AI-generated fake hotel reviews balanced across ten languages. Using this data, we performed extensive linguistic data analyses to gain insight into how AI fake hotel reviews differ from real hotel reviews. Finally, we explored the effectiveness of several models for deception detection in hotel reviews across three main dimensions: sentiment, location, and language. Despite the difficulty humans have in distinguishing between real hotel reviews and those generated by LLMs, we discovered that these posts have noticeable differences in style, structure, and semantics and that, even with little data, fine-tuned models accurately detect deceptive reviews across different languages. Our dataset is available for training and analyzing other models, and it can be accessed alongside our generation and classification models at https://anonymous.4open.science/r/hotel_reviews_deception.

Limitations

Multilingual Models' Limitations. When analyzing data across ten different languages, we encounter significant challenges in identifying computational models and tools that can be universally applied. In Table 2, we cannot find the LIWC categories required for analytic writing formula for the following languages: *German, Italian, Romanian, Korean, Russian, Turkish*. Additionally, the `textdescriptives` library does not currently support readability metrics for *Chinese*, *Korean*, and *Turkish*. This highlights the limitations of computational linguistic methods, which are currently predominantly English-focused. This, in turn, restricts the potential for research on multilingual data.

Data Generated with Closed-Source Model. We use GPT-4 to generate the hotel reviews, which is not an open-source LLM. We recommend future work to generate more data using open-source models like Mistral. At the same time, we are publicly releasing the data generated with GPT-4, so that others can also build on this dataset. We chose this model due to its SOTA performance and worldwide accessibility.

Ethical Considerations

We strongly oppose using our research findings and data to generate multilingual reviews to deceive consumers into believing they are human-authored. These unethical practices compromise the credibility of online review platforms and erode consumer trust. Instead, we advocate for transparency and authenticity in the digital marketplace. We have developed a multilingual deception detection model to combat the proliferation of fake reviews generated by bots. This model employs advanced algorithms to meticulously analyze linguistic nuances and syntactic structures, enabling the accurate differentiation between multilingual reviews created by language models and those written by human users. By providing this tool, we aim to empower businesses and online platforms to maintain ethical standards, protect consumers from deceptive practices, and foster a more trustworthy and reliable digital environment.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Duygu Altinok. 2023. A diverse set of freely available linguistic resources for Turkish. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13739–13750, Toronto, Canada. Association for Computational Linguistics.

Wissam Antoun, Virginie Mouilleron, Benoît Sagot, and Djamé Seddah. 2023. Towards a robust detection of language model generated text: Is chatgpt that easy to detect? *arXiv preprint arXiv:2306.05871*.

Hossein Bahak, Farzaneh Taheri, Zahra Zojaji, and Arefeh Kazemi. 2023. Evaluating chatgpt as a question answering system: A comprehensive analysis and comparison with existing models. *arXiv preprint arXiv:2312.07592*.

Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. 2019. Real or fake? learning to discriminate machine from human generated text. *arXiv preprint arXiv:1906.03351*.

Brett A Becker, Paul Denny, James Finnie-Ansley, Andrew Luxton-Reilly, James Prather, and Eddie Antonio Santos. 2023. Programming is hard-or at least it used to be: Educational opportunities and challenges of ai code generation. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, pages 500–506.

Amrita Bhattacharjee and Huan Liu. 2024. Fighting fire with fire: can chatgpt detect ai-generated text? *ACM SIGKDD Explorations Newsletter*, 25(2):14–21.

Cindy K Chung and James W Pennebaker. 2008. Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of research in personality*, 42(1):96–132.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2021. Is gpt-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Annual Meeting of the Association for Computational Linguistics*.

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. Tweepfake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Salvatore Giorgi, David M. Markowitz, Nikita Soni, Vasudha Varadarajan, Siddharth Mangalik, and H. A. Schwartz. 2023. "i slept like a baby": Using human traits to characterize deceptive chatgpt and human text. In *IACT@SIGIR*.

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. *arXiv preprint arXiv:2301.07597*.

Mandy Guo, Zihang Dai, Denny Vrandecic, and Rami Al-Rfou. 2020. Wiki-40b: Multilingual language model dataset. In *LREC 2020*.

Lasse Hansen, Ludvig Renbo Olsen, and Kenneth Enevoldsen. 2023. Textdescriptives: A python package for calculating a large variety of metrics from text. *arXiv preprint arXiv:2301.02057*.

Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Radar: Robust ai-text detection via adversarial learning. *Advances in Neural Information Processing Systems*, 36:15077–15095.

Austin Y Hubner and Robert Bond. 2022. I am a scientist... ask me anything: Examining differences between male and female scientists participating in a reddit ama session. *Public Understanding of Science*, 31(4):458–472.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2019. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*.

Maurice Jakesch, Jeffrey T Hancock, and Mor Naaman. 2023. Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. 2020. Automatic detection of machine generated text: A critical survey. *arXiv preprint arXiv:2011.01314*.

Marcia K Johnson and Carol L Raye. 1981. Reality monitoring. *Psychological review*, 88(1):67.

Kayla N Jordan, Joanna Sterling, James W Pennebaker, and Ryan L Boyd. 2019. Examining long-term trends in politics and culture through language of political leaders and cultural institutions. *Proceedings of the National Academy of Sciences*, 116(9):3476–3481.

Jason Kessler. 2017. Scattertext: a browser-based tool for visualizing how corpora differ. In *Proceedings of ACL 2017, System Demonstrations*, pages 85–90, Vancouver, Canada. Association for Computational Linguistics.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023a. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. 2023b. On the reliability of watermarks for large language models. *arXiv preprint arXiv:2306.04634*.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2024. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21258–21266.

David D. Lewis. 1998. Naive (bayes) at forty: The independence assumption in information retrieval. In *European Conference on Machine Learning*.

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. *ArXiv*, abs/1907.11692.

David M Markowitz. 2020. Putting your best pet forward: Language patterns of persuasion in online pet advertisements. *Journal of Applied Social Psychology*, 50(3):160–173.

David M Markowitz. 2023. Instrumental goal activation increases online petition support across languages. *Journal of Personality and Social Psychology*, 124(6):1133.

David M Markowitz and Jeffrey T Hancock. 2016. Linguistic obfuscation in fraudulent science. *Journal of Language and Social Psychology*, 35(4):435–445.

David M Markowitz and Jeffrey T Hancock. 2024. Generative ai are more truth-biased than humans: A replication and extension of core truth-default theory principles. *Journal of Language and Social Psychology*, 43(2):261–267.

David M Markowitz, Jeffrey T Hancock, and Jeremy N Bailenson. 2024. Linguistic markers of inherently false ai communication and intentionally false human communication: Evidence from hotel reviews. *Journal of Language and Social Psychology*, 43(1):63–82.

Tabea Meier, Ryan L Boyd, Matthias R Mehl, Anne Milek, James W Pennebaker, Mike Martin, Markus Wolf, and Andrea B Horn. 2020. Stereotyping in the digital age: Male language is "ingenious", female language is "beautiful"–and popular. *PloS one*, 15(12):e0243637.

Yisroel Mirsky, Ambra Demontis, Jaidip Kotak, Ram Shankar, Deng Gelei, Liu Yang, Xiangyu Zhang, Maura Pintor, Wenke Lee, Yuval Elovici, et al. 2023. The threat of offensive ai to organizations. *Computers & Security*, 124:103006.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*.

Artidoro Pagnoni, Martin Graciarena, and Yulia Tsvetkov. 2022. Threat scenarios and best practices to detect neural fake news. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1233–1249.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

James W. Pennebaker, Roger John Booth, and Martha E. Francis. 2007. Linguistic inquiry and word count (LIWC2007).

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate G. Blackburn. 2015. The development and psychometric properties of LIWC2015.

James W Pennebaker, Cindy K Chung, Joey Frazee, Gary M Lavergne, and David I Beaver. 2014. When small words foretell academic success: The case of college admissions essays. *PloS one*, 9(12):e115844.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

Vinu Sankar Sadasivan, Aounon Kumar, S. Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *ArXiv*, abs/2303.11156.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Chris Stokel-Walker. 2022. Ai bot chatgpt writes smart essays-should academics worry? *Nature*.

Student. 1908. The probable error of a mean. *Biometrika*, pages 1–25.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205*.

Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Sergey Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. 2024. Intrinsic dimension estimation for robust detection of ai-generated texts. *Advances in Neural Information Processing Systems*, 36.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Alham Fikri Aji, et al. 2023. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. *arXiv preprint arXiv:2305.14902*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F. Wong, and Lidia S. Chao. 2023a. A survey on llm-generated text detection: Necessity, methods, and future directions. *ArXiv*, abs/2310.14724.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F Wong, and Lidia S Chao. 2023b. A survey on llm-gernerated text detection: Necessity, methods, and future directions. *arXiv preprint arXiv:2310.14724*.

Xiao Yu, Yuang Qi, Kejiang Chen, Guoqiang Chen, Xi Yang, Pengyuan Zhu, Weiming Zhang, and Nenghai Yu. 2023. Gpt paternity test: Gpt generated text detection with gpt genetic inheritance. *arXiv preprint arXiv:2305.12519*.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *27th International Conference on Intelligent User Interfaces*, pages 841–852.

Appendix

A More about the Dataset

A.1 Real Hotel Reviews

Language We automatically web-crawl Booking.com for each of the ten languages by replacing the placeholder with the corresponding language abbreviation in the base URL: `https://www.booking.com/index.{language}.html`. For example, data in Turkish can be accessed via `https://www.booking.com/index.tr.html`.

To automate the web browsing process and make the data collection process more efficient, we use Selenium[10]. However, we observe that even if we browse in specific language settings, hotel reviews may still be in a mix of different languages. Therefore, we use the language filter bar to select the language we specify, which is also automated by using Selenium in the data collection process (Figure 5).

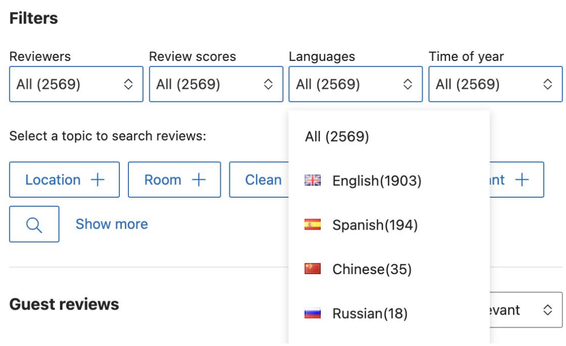

Figure 5: We use the language filter bar to select the language we specify, which is also automated by using Selenium in the data collection process.

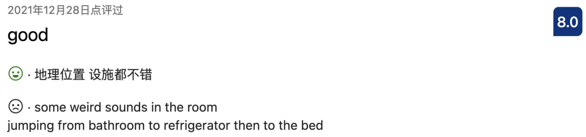

Figure 6: An example of a *Chinese* review where the "downside" part of the review is in *English*.

A.1.1 Data Quality

Automatic and manual language filtering. First, we filter out reviews based on length and language, that contain less than three tokens and are of a different language than the one crawled for. We use the `spaCy` library[11] to tokenize *Chinese* and *Korean* reviews, and `nltk`[12] for the rest of the languages. Next, we automatically verify the language of each review using the `langdetect` library.[13] We find several *Chinese* reviews were written in a combination of *Chinese* and another language, mostly *English*. Specifically, the "upside" or "downside" review may be in a different language, most commonly *English*, as seen in Appendix Figure 7. We choose to keep these reviews as they reflect the most realistic *Chinese* reviews written by people. Additionally, after a few manual language checks, we find that some *Chinese* reviews are incorrectly classified as *Korean*, and therefore chose to check all of them manually.

Detect fake human reviews. According to Booking.com, only tourists who have stayed in the hotel they booked or have visited the hotel but did not stay there can leave a review of the accommodation within three months of checking out. Additionally, Booking.com uses a combination of specialized personnel and an automated system to detect and remove fake reviews.

[10] https://www.selenium.dev/
[11] https://spacy.io/models
[12] https://www.nltk.org/
[13] https://pypi.org/project/langdetect/

A.2 LLM-generated Hotel Reviews

User Prompt Guidelines

> *Please write 4 prompts, 2 in English and 2 in {language}, asking GPT to write a hotel review. Each prompt should contain the following:*
>
> - *sentiment of the review: one positive and one negative for each: English and {language})*
> - *language of the review, translated in {language} for the {language} prompts*
> - *city of the hotel, translated in {language} for the {language} prompts*
> - *hotel name; just include it in the prompt as a placeholder: "hotel X", so you just translate the word hotel for the {language} prompts*

Challenges solved with few-shot prompting. The prompt design process is complex because we aim to generate reviews that capture the format of the real hotel reviews. A real review contains two sections: the "upside" and the "downside", each with corresponding sentiment. At the same time, the review's sentiment is given from the combined assessment of both the "upside" and the "downside". In addition, either the "upside" or "downside" could be null. When we include this specification in the instruction, we find that the reviews generated

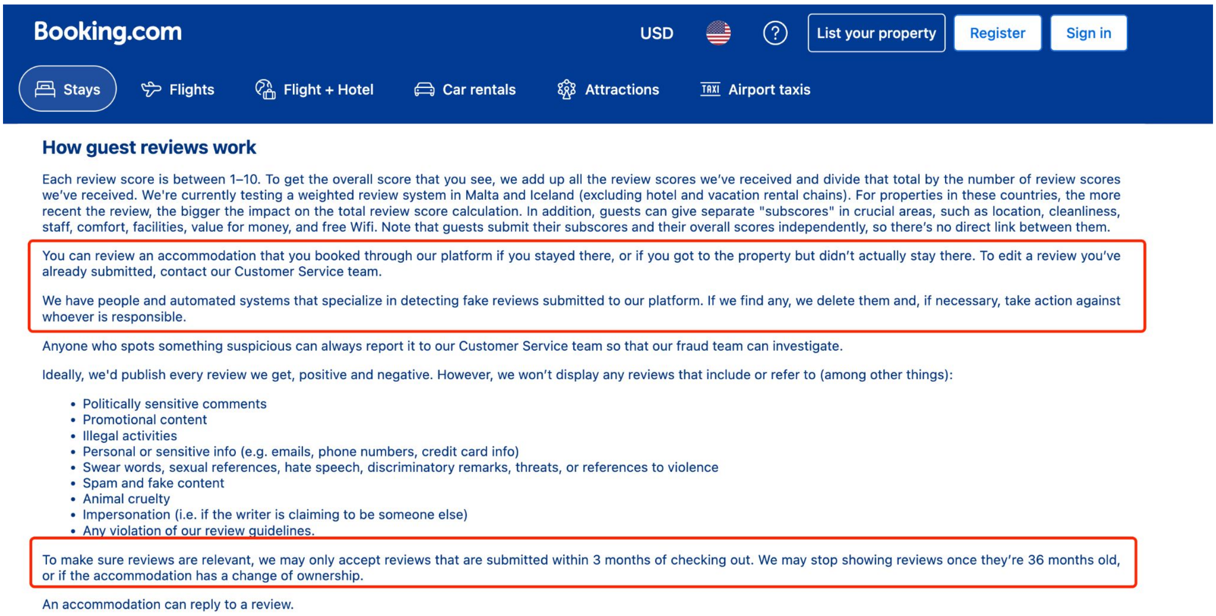

Figure 7: Specifications from Booking.com regarding checking for fake reviews

with a positive sentiment predominantly contain content in the "upside" section, and no content in the "downside", and conversely for reviews with negative sentiment. We solve this issue by using a few-shot prompting approach, as mentioned above.

B Data Analysis

Word frequency. We compute the most frequent n-grams across each language after pre-processing the text, which is shown in Table 9.

Topic Modeling

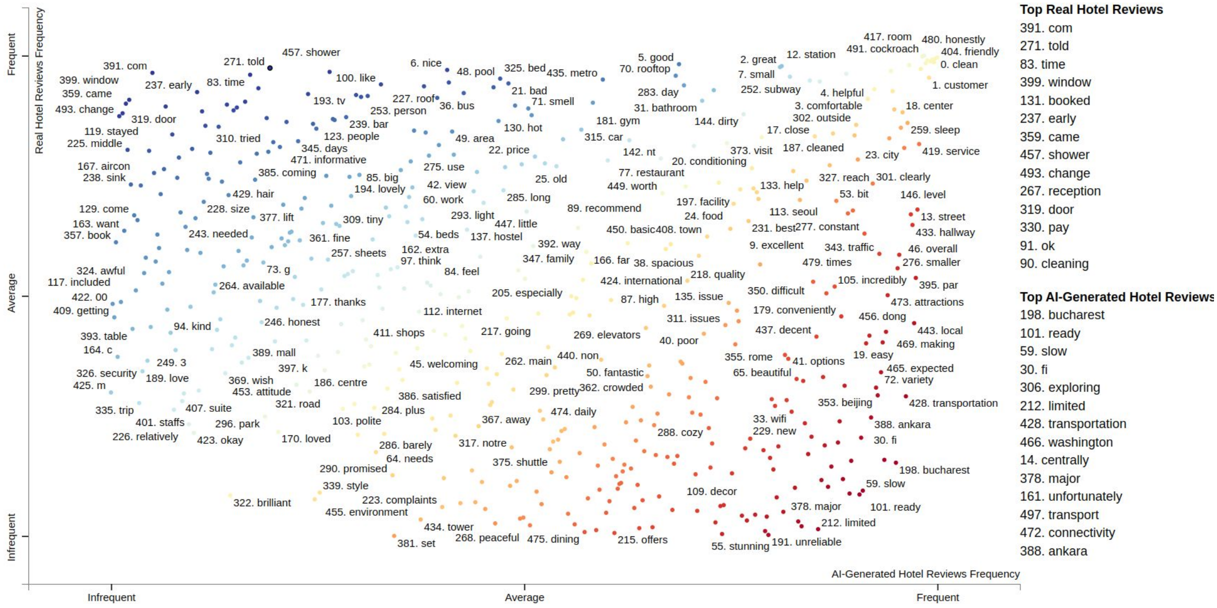

Figure 8: Visualization of **topics** used in the real and LLM-generated English reviews. Points are colored red or blue based on the association of their corresponding topics with AI-generated or real hotel reviews. The most associated topics are listed under **Top AI-Generated** and **Top Real** headings. Interactive version: `https://anonymous.4open.science/r/hotel_reviews_deception`.

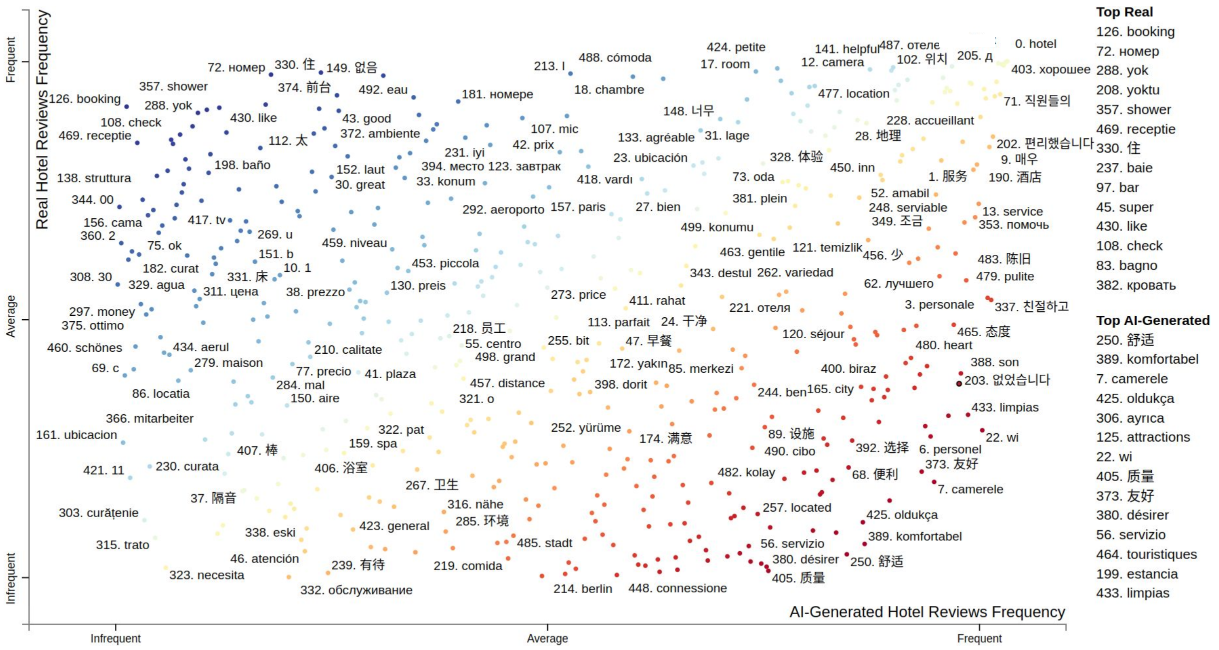

Figure 9: Visualization of **topics** used in the real and LLM-generated reviews across all languages. Points are colored red or blue based on the association of their corresponding topics with AI-generated or real hotel reviews. The most associated topics are listed under **Top AI-Generated** and **Top Real** headings. Interactive version: `https://anonymous.4open.science/r/hotel_reviews_deception`.

Lang	Prompt/ User Message
Ch	+ 请模仿人类用*L*为*C*的*H*酒店写一条正面评价 T : please write a positive review using L for hotel H in C that would mimic a human writing a hotel review - 请模仿人类用*L*为*C*的*H*酒店写一条负面评价 T : please write a negative review using L for hotel H in C that would mimic a human writing a hotel review
Fr	+ *Écrivez un avis positif sur l'hôtel H en C. Veuillez écrire en L e parler des différents aspects de votre séjour.* T : Write a positive hotel review for hotel H in C. Please write it in L and talk about different aspects of your stay. - *Écrivez un avis négatif sur l'hôtel H en C. Veuillez écrire en L e parler des différents aspects de votre séjour.* T : Write a negative hotel review for hotel H in C. Please write it in L and talk about different aspects of your stay.
Ge	+ *Du bist ein Tourist, und du hast in hotel H in der Stadt C übernachtet. Das hotel hat dich richtig gefallen. Schreib ein Positives Review auf L über dein Stay, was fand besonderes gut and ob du das empfehlen würdest.* T : You are smart and helpful assistant. Your goal is to write a positive and realistic review for the hotel H in the language L, where you stayed in the city C. Make sure to mention why you enjoyed your stay and list all the positive features of the hotel. - *Du bist ein Tourist, und du hast in hotel H in der Stadt C übernachtet. Das hotel hat dich gar nicht gefallen. Schreib ein Negatives Review auf L über dein Stay, was fandest du besonderes schlecht and warum du das Hotel nicht empfehlen würdest.* T : You are smart and helpful assistant. Your goal is to write a negative and realistic review for the hotel H in the language L, where you stayed in the city C. Make sure to mention why you disliked your stay and list all the negative features of the hotel.
It	+ *Scrivi una recensione positiva in L per l'albergo H di C. Parla di almeno tre aspetti diversi del tuo soggiorno.* T : Write a positive review in L for hotel H in C. Talk about at least three different aspects of your stay. - *Scrivi una recensione negativa in L per l'albergo H di C. Parla di almeno tre aspetti diversi del tuo soggiorno.* T : Write a negative review in L for hotel H in C. Talk about at least three different aspects of your stay.
Ko	+ *지난주에 갔던 C의 호텔 H이 너무 맘에 들었다고 L로 리뷰좀 남겨줄래?* T : I really enjoyed my stay at Hotel H in C last week. Can you write a review for me in L? - *L로 C에 있는 호텔 H이 너무 별로였다고 평점 좀 남겨줘.* T : Can you write a review that the hotel H that we stayed at in C last week was terrible? Can you write it in L?
Ro	+ *Scrie un comentariu pozitiv in limba L pentru hotelul H din orașul C. Scrie la fel ca un român care a vizitat hotelul și a lăsat un comentariu.* T : Write a negative sentiment review in L language for the hotel H from C. - *Scrie un comentariu negativ in limba L pentru hotelul H din orașul C. Scrie la fel ca un român care a vizitat hotelul și a lăsat un comentariu.* T : Write a positive sentiment review in L language for the hotel H from C.
Ru	+ Представь что ты турист, и тебе очень понравилось твое пребывание в отеле H в городе C. Напиши реалистичный отзыв на L языке об этом отеле. Упомяни все черты которые тебе понравились и почему ты порекомендовала бы этот отель другим туристам. T : You are a tourist and you really enjoyed staying in the hotel H in the city C. Write a simple hotel review in language L, where you mention all the positive features of the hotel and how much you liked them. - Представь что ты турист, и тебе очень не понравился отель H в городе C. Напиши реалистичный отзыв на L языке об этом отеле. Упомяни все черты которые тебе не понравились и почему бы ты не порекомендовала этот отель другим туристам. T : You are a tourist and you seriously dislike your stay in the hotel H in the city C. Write a simple hotel review in language L, where you mention all the things you disliked, and why you wouldn't recommend this hotel to the other tourists.
Sp	+ *¿Podés escribir un comentario positivo en L sobre el hotel H de C?* T : Can you please write a positive review in L for the hotel H located in C? - *Escribí un comentario negativo en L sobre un hotel llamado H que está en C.* T : Write a review that's negative in sentiment in L language for a hotel named H that's located in C.
Tu	+ *C H için olumlu yorumu L olarak yaz.* T : Write a review in L about a negative experience staying in the H hotel in C. - *C H de kötü bir zaman geçirdiğini duşun. L olarak H hakkında olumsuz yorum yaz.* T : Imagine that you had an enjoyable stay at the hotel H while visiting C. Write a review in L to describe your experience to others who are considering booking a stay at H.

Table 6: Prompts/ System messages in each language, together with their English translation marked with T.

Language	Real Review	Generated Review
Chinese	+ 早餐丰富，服务态度很好，房间很整洁，离地铁站很近。优秀 - 交通方便价格性价比不错; 设施有些陈旧可能因为疫情的缘故卫生各个地方显得很陈旧的污渍打扫看着不是那么上心早上在酒店大堂看见任何一个服务员均很冷漠没有打招呼的习惯希望能改进	+ 这家酒店的位置极佳，位于市中心，靠近主要的旅游景点。酒店的装修新颖，服务人员非常专业和友好。早餐也很美味，种类丰富。 - 该酒店地理位置优越，交通便利。我订的房间缺乏维护，房间内设施感觉比较陈旧，尤其是地毯，看上去非常脏旧，且有种异味。另外，酒店早餐品种不多，食品口味偏弱。最不满的是酒店的服务态度，服务员的服务意识强度不足，没有解决问题的积极性和主动性，完全不符合四星级酒店的标准。
English	+ *Nicely furnished room and nicely decorated lobby. Room service is affordable and the receptionists, especially Mr. Omar, are usually eager to help with any queries.* - *The location was very central and the staff was nice and helped us.; The room smelled like cigarettes and there was mold in the bathroom*	+ *The staff was extremely helpful and accommodating. Clean and well-furnished rooms. Central location with easy access to public transport.; The breakfast offerings could be more varied.* - *Nothing. Horrible experience.; Bad customer service, rooms were not clean and the food was below average.*
French	+ *Pas besoin de tourner pendant un moment pour trouver une place de Parking, un membre de l'hôtel se charge de prendre votre voiture; En soirée la rue n'est pas très familiale* - *Bon emplacement dans Ankara. Personnel très agréable.; Petit déjeuner moyen. Fuites dans la salle de bain. N'est pas du niveau d'un hôtel 5 étoiles.*	+ *L'hôtel est juste à côte de l'aéroport, parfait pour les vols tôt le matin. Les chambres étaient propres et confortables.; La diversité de la nourriture pourrait être améliorée.* - *L'emplacement est bien, situé à proximité du centre-ville d'Ankara; Le manque de propreté est notable. La nourriture n'est pas de bonne qualité et le service à la clientèle laisse à désirer. Les chambres sont bruyantes et mal isolées, et les meubles sont vieux et usés.*
German	+ *Die Lage ist super zentral! Zu Fuß nur 10 Minuten vom Kızılay Platz entfernt. Die Zimmer waren sehr sauber und sehr geschmackvoll eingerichtet. Das Frühstück war sehr reichhaltig und lecker. Das Personal ist sehr freundlich und hilfsbereit.; Keine* - *Das hotel ist einfach viel zu alt. Ehrlich gesagt ist es 3 sterne; Nicht sauber habe nicht im hotel duschen können. Wir hatten gesamt zwei zimmer beide schlecht sauber gemacht worden*	+ *Das Personal war sehr freundlich und das Zimmer war sehr sauber. Die Lage ist fantastisch, nahe an vielen Sehenswürdigkeiten. Das Frühstück hatte eine gute Auswahl an Speisen.* - *Die Lage des Hotels ist sehr zentral; Der Service war sehr schlecht und die Sauberkeit ließ zu wünschen übrig.*
Italian	+ *Struttura bella e pulita. Vicinissimo alla metro e centro città raggiungibile a piedi. Letto e cuscini comodissimi.; Il bagno turco aveva un odore strano.. muffa?* - *Staff disponibile e cordiale; La camera doveva essere doppia matrimoniale, ma il letto effettivamente è una piazza e mezzo, la moquet era macchiata.*	+ *Camera molto pulita e silenziosa, personale professionale e disponibile. La colazione era abbondante e deliziosa. Posizione eccellente nel centro della città.* - *La posizione è centralissima, vicino a molti negozi e ristoranti.; Stanze poco pulite e personale maleducato. Colazione molto scarsa e Wi-Fi praticamente inesistente.*
Korean	+ *직원들이 자기일처럼 성실히 도와주었으며 거실,방마다있는 에어컨도 방성능이 좋았으며,조명도 밝고 시설이 깨끗해서 편안했습니다; 세탁기는 있는데 드라이어나 건조대가 없어 옷 말리기가 불편했습니다* - *영등포역에서 가까워요. 침구가 깨끗합니다. 방은 깨끗; 전철이 있어서 시끄럽다는 후기가 있어 씨티뷰로 신청했지만 마찬가지로 소음이 심했습니다. 실내가 좁아 불편. 주차요금을 내야합니다. 화장실이 문이 없습니다*	+ *뉴 파크 호텔'은 매우 좋았습니다. 직원들이 친절하였고, 호텔 시설도 깨끗했습니다. 특히 아침식사가 훌륭했어요.; 방음이 잘 되지 않아 약간 소음이 있었습니다. 하지만 많이 거슬리지는 않았어요.* - *파리 중심부에 위치해 있어서 접근성이 좋았습니다.;객실이 너무 좁아서 실망했고, 침구가 깨끗하지 않았습니다. 또한, 냉난방 시스템이 제대로 동작하지 않아서 매우 불편했습니다.*

Table 7: Random review samples with positive (+) and negative (-) sentiments across the 10 languages.

Language	Real Review	Generated Review
Romanian	+ *Gazde foarte amabile, primitoare, camera mare, curată, paturi confortabile, parcare proprie în fața hotelului, micul dejun, bufet, diversificat.* - *E ffff cald in camere! Și aparatele stricate, camerele mici!!*	+ *Servicii excelente, camere curate și spațioase. Spa-ul și piscina sunt minunate pentru relaxare."; Nu am avut nicio problemă sau dezamăgire.* - *Amplasarea hotelului este convenabilă, relativ aproape de majoritatea obiectivelor turistice.; Confortul camerei a fost sub orice critica, patul a fost foarte inconfortabil, iar curăţenia a fost precară. De asemenea, personalul a fost nepoliticos si neajutorător.*
Russian	+ Шикарный завтрак, удобное расположение, доброжелательный персонал, согласовали ранний заезд; Немного грязное окно и тусклое освещение, но это мелочи - метро рядом , санлитунь рядом.; персонал за стойками регистрации, это просто треш, мало того, что цены в чеке не совпадали с ценой на букинге, так еще и чекаут оформляли 50 минут! это дно товарищи!	+ Отель с удобным расположением, близко к основным достопримечательностям. Очень чистые и комфортные номера. Отзывчивый и вежливый персонал.; Сложно было найти русскоговорящий персонал. Завтраки могли бы быть более разнообразными. - Нет положительных сторон.; Большая часть персонала не говорит по-английски, что особенно усложнило процесс общения. На завтраке был ограниченный выбор блюд, а в номерах отличался сильный запах курения.
Spanish	+ *Me gustó todo, super bonito, limpio, cómodo y en una zona muy animada, con un desayuno buffet genial; Nada* - *Hotel con bien trato, y muy céntrico, limpio y comodo; Discoteca justo debajo de la habitación muchísimo ruido y música disco mucho volumen hasta altas horas. No se puede descansar*	+ *Un hotel muy bueno. Estaba muy limpio y moderno. En el corazón de la ciudad, cerca de las tiendas y restaurantes. El personal fue amigable y la comida era buena.; La única queja que tenía era que el wifi en mi habitación era un poco lento.* - *Buena ubicación central y la habitación estaba limpia; El desayuno era muy básico y la presión del agua en la ducha era bastante baja*
Turkish	+ *Kahvaltısı çok güzeldi. Tam bir Fransız kahvaltısıydı. Personel çok güler yüzlüydü. Otel çok merkezi bir konumdaydı. Tekrar Paris'e gelirsem tercihim yine buradan yana olur.* - *Bookingde sigarasız yazmasına rağmen ortak havalandırmadan sigara dumanı geliyordu wi-fi çalışmıyordu*	+ *Otelin konumu ve erişilebilirliği mükemmel. Odalar temiz, konforlu ve fonksiyonel.; Oda sıcaklık ayarları biraz daha iyi olabilirdi.* - *Otelin konumu iyi.; Oda temizliği yetersizdi, yemekler çok tuzluydu ve personel pek yardımcı olmadı. Bu nedenlerden dolayı diğer turistlere bu oteli önermiyorum.*

Table 8: Random review samples with positive (+) and negative (-) sentiments across the 10 languages.

	uni-gram		bi-gram		tri-gram	
Lang	**Real**	**Gen**	**Real**	**Gen**	**Real**	**Gen**
	酒店 房间 位置	酒店 位置 服务	(地理, 位置) (工作, 人员) (位置, 不错)	(地理, 位置) (酒店, 位置) (服务, 人员)	(地理, 位置, 不错) (地理, 位置, 优越) (员工, 服务, 态度)	(酒店, 地理, 位置) (服务, 人员, 态度) (地理, 位置, 优越)
	room location hotel	room hotel location	(great, location) (room, clean) (location, good)	(room, clean) (staff, friendly) (customer, service)	(staff, friendly, helpful) (staff, nice, helpful) (staff, speak, english)	(room, clean, comfortable) (leave, lot, desire) (staff, friendly, helpful)
	chambre petit hôtel	chambre hôtel personnel	(petit, déjeuner) (salle, bain) (bien, situer)	(petit, déjeuner) (chambre, propre) (chambre, spacieux)	(rapport, qualité, prix) (hôtel, bien, situer) (bon, petit, déjeuner)	(petit, déjeuner, varier) (chambre, propre, confortable) (chambre, spacieux, propre)
	Zimmer Lage Hotel	Hotel Zimmer Lage	(freundlich, Personal) (Personal, freundlich) (zentral, Lage)	(Lage, Hotel) (freundlich, hilfsbereit) (Personal, freundlich)	(Personal, freundlich, hilfsbereit) (Lage, freundlich, Personal) (Personal, super, freundlich)	(Personal, freundlich, hilfsbereit) (lassen, wünschen, übrig) (Lage, Hotel, zentral)
	posizione camera colazione	posizione hotel camera	(ottimo, posizione) (posizione, ottimo) (personale, gentile)	(camera, pulito) (posizione, centrale) (posizione, hotel)	(rapporto, qualità, prezzo) (personale, gentile, disponibile) (ottimo, rapporto, qualità)	(personale, cordiale, disponibile) (camera, pulito, confortevole) (personale, gentile, disponibile)
	너무 좋았습니다 위치가	매우 또한 호텔	(바로, 앞에) (너무, 좋았어요) (위치가, 너무)	(호텔의, 위치는) (방음이, 되지) (또한, 직원들의)	(위치가, 너무, 좋았고) (바로, 앞에, 있어서) (좋음, 직원, 친절함)	(방음이, 되지, 않아) (방음이, 되지, 않아서) (친절하고, 도움이, 되었습니다)
	cameră mic hotel	cameră personal hotel	(mic, dejun) (personal, amabil) (cameră, mic)	(mic, dejun) (personal, amabil) (personal, prietenos)	(mic, dejun, bun) (mic, dejun, bogat) (mic, dejun, ok)	(cameră, curat, confortabil) (aproape, centru, oraș) (mic, dejun, putea)
	номер отель завтрак	отель номер персонал	(постельный, бельё) (горячий, вода) (приветливый, персонал)	(центр, город) (желать, хороший) (wi, fi)	(соотношение, цена, качество) (постельный, бельё, полотенце) (персонал, говорить, английский)	(оставлять, желать, хороший) (дружелюбный, готовый, помочь) (бесплатный, wi, fi)
	habitación ubicación hotel	habitación hotel ubicación	(personal, amable) (habitación, pequeño) (aire, acondicionado)	(personal, amable) (dispuesto, ayudar) (ubicación, hotel)	(relación, calidad, precio) (personal, amable, habitación) (cerca, torre, eiffel)	(amable, dispuesto, ayudar) (personal, amable, dispuesto) (personal, amable, servicial)
	bir konum od	oda otel olmak	(konum, iyi) (gülmek, yüz) (od, Küçük)	(otel, konum) (oda, temiz) (yardımcı, olmak)	(çalışmak, gülmek, yüz) (personel, gülmek, yüz) (yer, yürümek, mesafe)	(personel, son, derece) (oda, te, konfor) (oda, temiz, yeter)

Table 9: Most frequent n-grams for real and generated data, for each review language. Country flag according to review location.