**Fake-Review Generation**

Jitea Gabriel-Octavian, Rus Alexandru

**Introduction**

As we have discussed in the previous milestones, customer reviews represent a huge part of online shopping and are crucial for the success or fail of a business that tries to sell services or products online. However, this increasing reliance on reviews has also made them a target for manipulation. Fake reviews—both positive and negative—can distort product reputations, mislead consumers, and erode trust in digital platforms. As a response, the ability to detect and analyze fake reviews has become a vital task in maintaining content integrity.

This part of the project focuses on the **generation of high-quality fake reviews** for the purpose of training and evaluating machine learning models capable of detecting fake content. Our goal is to construct a **realistic, balanced, and diverse dataset** combining both genuine and synthetic reviews. This will serve as a foundation for building robust classifiers and uncovering different patterns of deception used in review manipulation.

We begin with a base of **911,721 real reviews** and aim to introduce an additional **12,000 fake reviews.** Using the approaches we will describe, we aim to simulate a wide spectrum of fake review styles, from blatantly fabricated to subtly manipulated. This diversity is essential for training models that generalize well to real-world scenarios. Ultimately, this project supports the development of detection systems that help safeguard platforms from misinformation and manipulation.

**Step 1. AI-Generated Reviews**

To simulate synthetic fraudulent review patterns, we first leverage **Large Language Models (LLMs)** such as ChatGPT to generate high-quality fake reviews. These AI-generated reviews form the main part of our synthetic dataset and provide diverse, scalable, and controllable examples of deceptive review content. This step is crucial for building a robust classifier capable of detecting subtle and complex forms of fake reviews.

**Objective**

Generate **10,000 synthetic fake reviews** that mimic deceptive content using AI to complement our set of **911,721 real reviews**. This data will support the development of a binary (real vs. fake) or multiclass (real, AI-generated, heuristic, augmented) classifier.

* 1. **Model Selection**

We utilize state-of-the-art LLM -> **OpenAI's GPT-4/ChatGPT.**

* 1. **Prompt Engineering**

To create diverse and realistic fake reviews, prompts are carefully crafted to reflect:

* Positive vs. negative sentiment
* Varying lengths (short blurbs to multi-paragraph reviews)
* Multiple domains (electronics, fashion, home goods, etc.)
* Tones (professional, casual, overly enthusiastic, suspiciously generic)

| **#** | **Prompts** |
| --- | --- |
| 1 | Generează un review fals scurt, suspicios de pozitiv, pentru un telefon mobil. |
| 2 | Scrie un review vag și prea entuziast pentru un uscător de rufe. |
| 3 | Creează un review fake pentru o mașină de spălat care nu are sens logic. |
| 4 | Scrie un review fals, scurt și fără detalii utile, pentru un laptop. |
| 5 | Generează un review fake foarte general pentru o carte. |
| 6 | Creează un review fals care exagerează calitatea unei huse de telefon. |
| 7 | Scrie un review fals pentru un blender care pare scris de un robot. |
| 8 | Generează un review fake pentru un televizor, cu afirmații contradictorii. |
| 9 | Scrie un review fals, cu greșeli gramaticale și entuziasm exagerat, pentru o pereche de căști. |
| 10 | Creează un review fals foarte vag, pentru un produs de baie. |

**These reviews have been generated very fast after selecting the proper prompts so this method is very scalable as it is easy to generate large amounts of data, pretty controllable as we can select the tone, domain and deception features while also having variety, models which can simulate multiple fake styles (exaggerated, vague, copied).**

Generated reviews will be stored in a structured format (.CSV) and integrated with:

* **Heuristic-based reviews** (Step 2)
* **Real reviews** (existing base)

This allows for downstream training, validation, and interpretability of fake review detection models.

**Step 2: Heuristic-Based Reviews**

***Overview***

In this step, we generate fake reviews by teaching a language model how real users write, and then using it to produce new—but deliberately artificial—reviews. The idea is to blend authentic linguistic patterns with deceptive content cues to create reviews that *look real* but *feel off*—just like the kind of manipulative content we want to detect.

Unlike direct prompting of large language models, this approach gives us more control and realism by training the model specifically on Romanian user reviews, generating reviews from template-like prompts that reflect common but shallow review styles and introducing heuristic markers of deception—such as vagueness, repetition, and excessive positivity

By doing this, we aim to create a set of synthetic reviews that are both highly plausible and intentionally suspicious.

***Step-by-Step***

1. **Training a Romanian Review Model**

We started with readerbench/RoGPT2-medium, a Romanian GPT-2 variant from HuggingFace. To make the model better understand the tone, vocabulary, and structure of genuine Romanian reviews, we finetuned it on a dataset of real reviews using the following setup:

* **Dataset**: romanian\_real\_reviews.txt
* **Max length**: 128 tokens
* **Epochs**: 2
* **Batch size**: 2



**Step 2: Generating Fake Reviews using the trained model**

Now that we had a model fluent in Romanian review style, we used it to generate **2,000 fake reviews** by giving it templated prompts like:

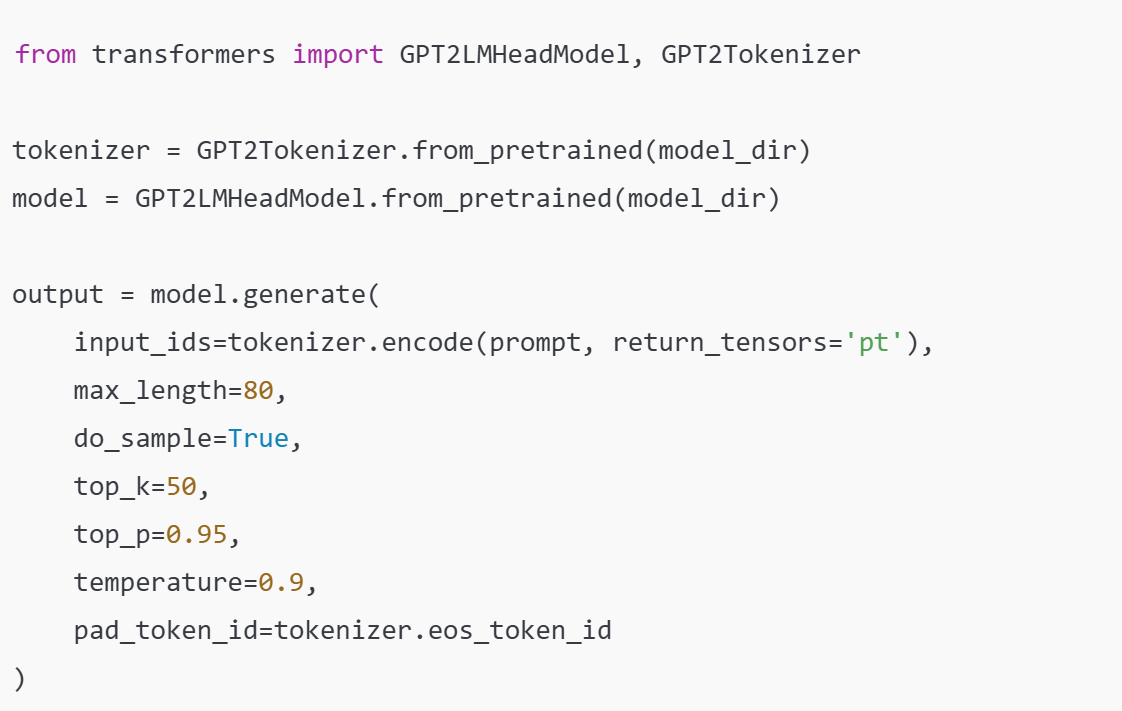
 “Am cumpărat acest produs de pe eMAG și...”

 “Sunt foarte mulțumit de achiziție pentru că...”

 “Pentru prețul plătit, consider că...”

 “Funcționează bine cu alte accesorii și...”

Sample Code:



These methods were able to give us the possibility to generate a dataset of our own as there is not a big selection of datasets containing fake reviews out there. This way, we can train a classifier which has the knowledge to predict a fake or real review and it can solve real issues for businesses as well as buyers.

**Step 3: Detection of Fake Reviews**

After generating the fake reviews, we sample real reviews to ensure two key aspects are maintained:

* They are approximately evenly distributed across the most popular product categories.
* They reflect a variety of review lengths.

For the detection of fake reviews, we utilized the pre-trained BERT model fine-tuned on our dataset. The text data was tokenized, converted to lower-case, stripped of punctuation, and filtered to remove stopwords. The dataset was divided into training (70%), validation (15%), and test (15%) sets. The training and validation sets were used to fine-tune the BERT model, while the test set was reserved for evaluation. We employed the BertTokenizer for tokenization, ensuring uniform input length by truncating or padding sequences to a maximum length of 512 tokens. The fine-tuning was conducted using the Hugging Face Trainer API. We used the following hyperparameters:

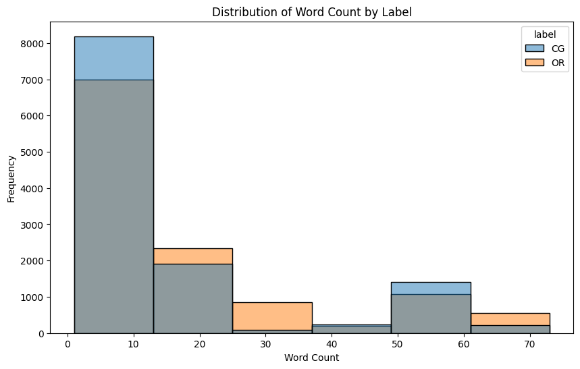
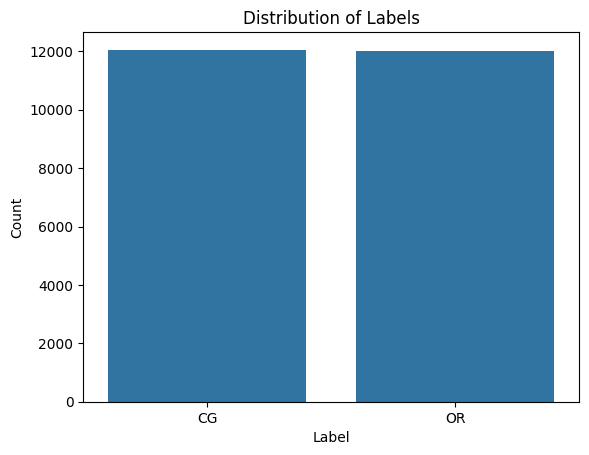
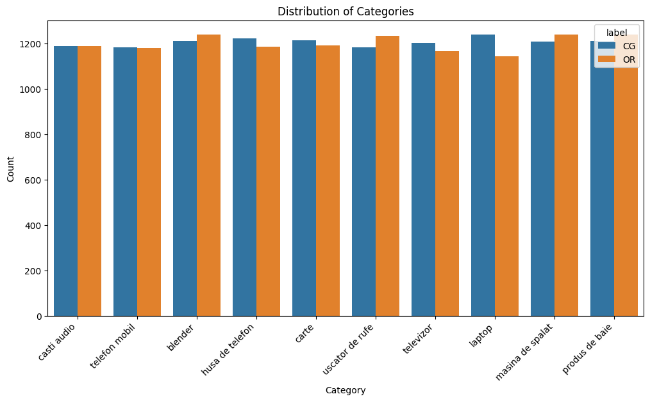
• Learning rate: 2 × 10−5

• Batch size: 16

• Number of epochs: 20

• Weight decay: 0.01

**Distribution of the data across categories, labels and word count:**



The training process saved the best model based on validation accuracy at the end of each epoch. The fine-tuned BERT model achieved a test accuracy of 0.983. This high accuracy demonstrates the effectiveness of BERT, though it comes with the drawback of requiring significant computational resources and extended training time. The improvement from the results on the previous dataset (accuracy of 0.946) suggests that the generated fake reviews are easier to distinguish from real ones, leaving room for further improvement in their creation.