Submission from News Summary Team:
Seida Basha, Lily Djami, Eric J. Herschbach, and Samuel Khodak

Requirements Document for News Summary Project

1.Task

1.1. Type

The task in this project is to develop a web app that summarizes English news (1-3 sentences per article) from various sources, as well as filter fake news. The NLP tasks in this project are to be done using the novel GPT-3 model.

1.2. Objective

The objective of this project is to help people who don't have much time, but want to stay up to date with the latest news and prevent the spread of misinformation.

1.3. Description

Every day the application will summarize news from the past 24 hours while also filtering out fake news.

1.4. Intended audience

The end product of this project is intended for the general audience, anyone who wants to keep up with the latest news.

1.5. Intended use

The user should be able to just open the application in the morning and receive a summary of news from that day (the past 24 hours).

1.6. Previous procedure

Previously, users would need to spend a longer time reading news articles from various news sites to keep in touch with the latest news.

1.7. Characteristics to be validated

The quality of summaries generated are to be tested using human testers. We evaluate the generated summaries based on the following criterias:

- Most keywords (or their synonyms) are included in the summary.
- The meaning of the article must not be changed in the summary.
- No false statements have been made in the summary.
- The generated text is contextually sound.

We will also test the user interface of our project to ensure its functionality.

1.8. Functional features

By the end of the project, the following features should be implemented:

- Summary generation
- Fetching news from the past 24 hours
- Fetching news only from a specific category (i.e. sports news, tech news).
- Summary generation using manual input

1.9. Availability

The web app should be available for users 24/7, on-demand. Internet connection must be available to the user during operation.

2.Test Object

2.1. Types

The test objects in this project are news articles (i.e. from CNN/BBC).

2.2. Sets

Each test set contains a different number of real and fake news.

3.Scene

3.1. News Definition

According to the Oxford dictionary, news can be defined as:
“a broadcast or published report of news.”
For this project, we focus on news articles taken from news websites such as CNN or BBC.

3.2. Summary

A good summary should reflect a news article in a shorter and more compact text than the original. It must contain all main aspects and events, so that it can be read without needing to read or know the original’s content. The summary should be shorter by at least half the length of the original in number of words.

4.Process Integration

4.1. Use of data

Data will be required to train and improve the GPT-3 model for the specific use case in this project.

4.2. Input and output

The input in this project is news articles, either fetched automatically from the URL or entered manually by copying and pasting the text of the article. The output is summaries as well as a reliability score of the news articles.

4.3. Dataset

For this project, we are using the following datasets:

- CNN, BBC, etc. (news articles by reliable sources)
- https://tagoyal.github.io/zeroshot-news-annotations.html (a similar project that provides datasets)
- https://huggingface.co/datasets/cnn_dailymail

4.4. Operating modes

The web app should have two modes of operation:

- Automatic: web app fetches news from the past 24 hours on set intervals and generates summaries from them.
- Manual input: user enters a news article manually and the web app generates a summary from it.

4.5. External interface requirement:

API to work with GPT-3

5.Human Machine Interface

5.1. Operating concept

As the intended application will be web-based, the website will serve as the graphical user interface. A sidebar will be used to navigate to different parts of the website. Interactive elements of the user interface will have tooltips to inform users of their uses. The website will be available in English.

6.Misc

6.1. Installation

Installation is not needed, as the intended app is a web app, which can be opened in any browser through the given URL.

6.2. Training

Special training for using the application is not needed, the user interface should be intuitive enough.

6.3. Documentation

A documentation of the web app will be available on the website.

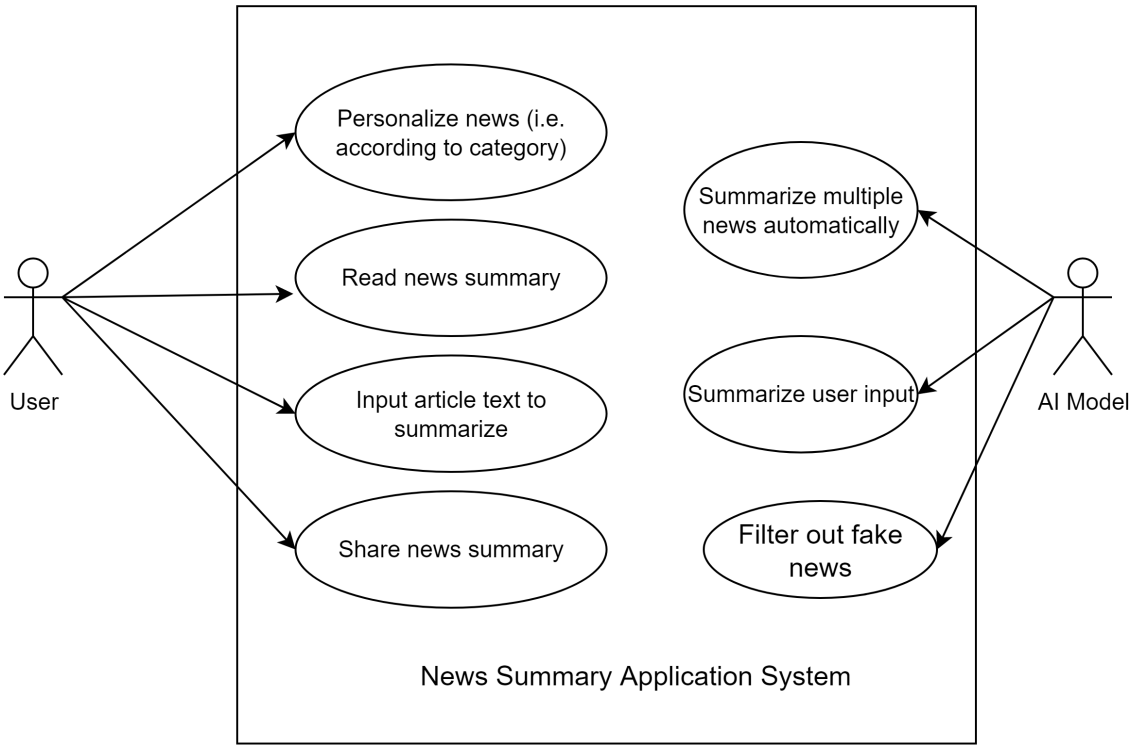
6.4. Maintenance

Maintaining this web application requires server maintenance, which will not be covered after the end of the project.

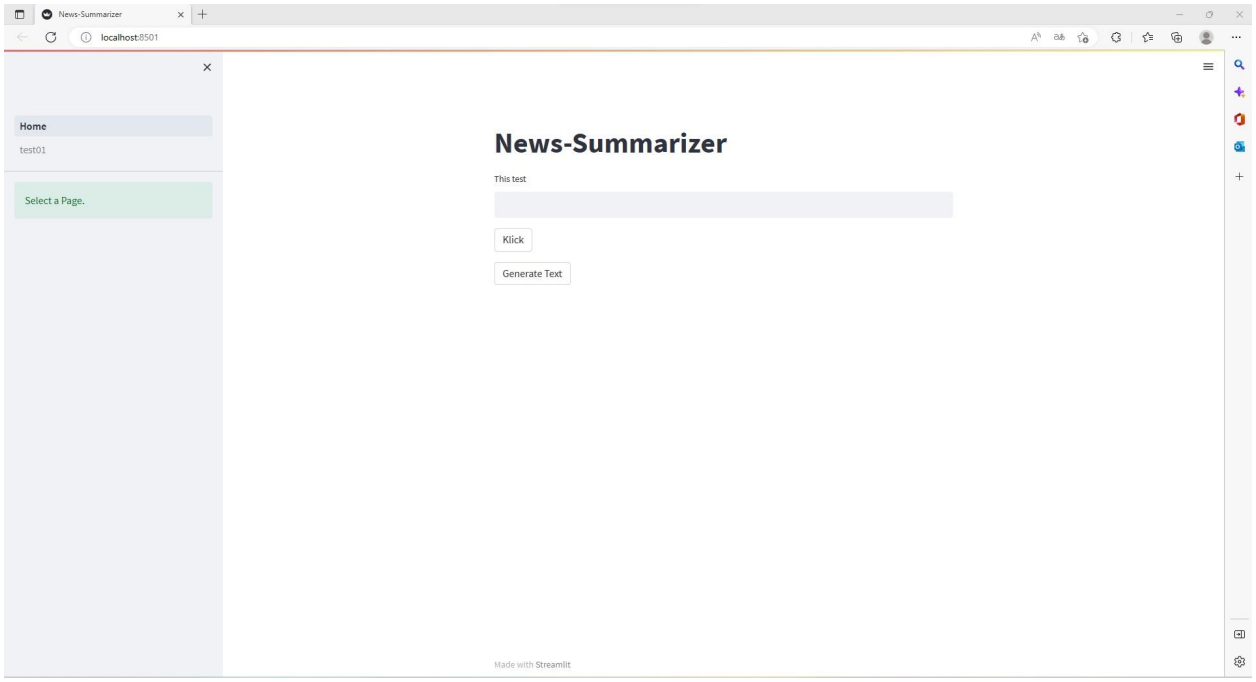
6.5. Stability monitoring

For the duration of the project, the API to GPT-3 will be checked monthly to make sure that it is working.

Use Case Diagram from the Perspective of the User



First working page



Design Documents

1 Implementation Details

1.1. Frontend

The Web app frontend will be designed with python/streamlit

There will be a:

- Homepage
    - Some random summarized news articles
- Multiple category pages
- A page where users can summarize their own news article
- Documentation on how this page works and how to use it
- Contact site
- Imprint

1.2 Backend

The backend consists of 2 parts, the GPT-3 summarizer and a news fetching mechanism.

ML Summarizer

We tested the GPT-3 TL;DR summarizer by OpenAI. The summary generated by this model works well, however the summary is not in the bullet point format that we wanted. Upon consultation with the ML team, it is not possible to fine-tune the GPT-3 model to suit our needs. Thus, we decided to use a custom model for our summarizer.

> An example of the GPT-3 TL;DR summarizer
>
> Article: Iran: fears grow of security crackdown in Zahedan as anti-regime protests persist
> Summary:
> Thousands of protesters are marching in the city of Zahedan, Iran to protest against the government. There has been an increased presence of security forces in the city, with 15 checkpoints set up and reports that protesters have been detained by security forces. Human rights activists fear that the situation may become dangerous

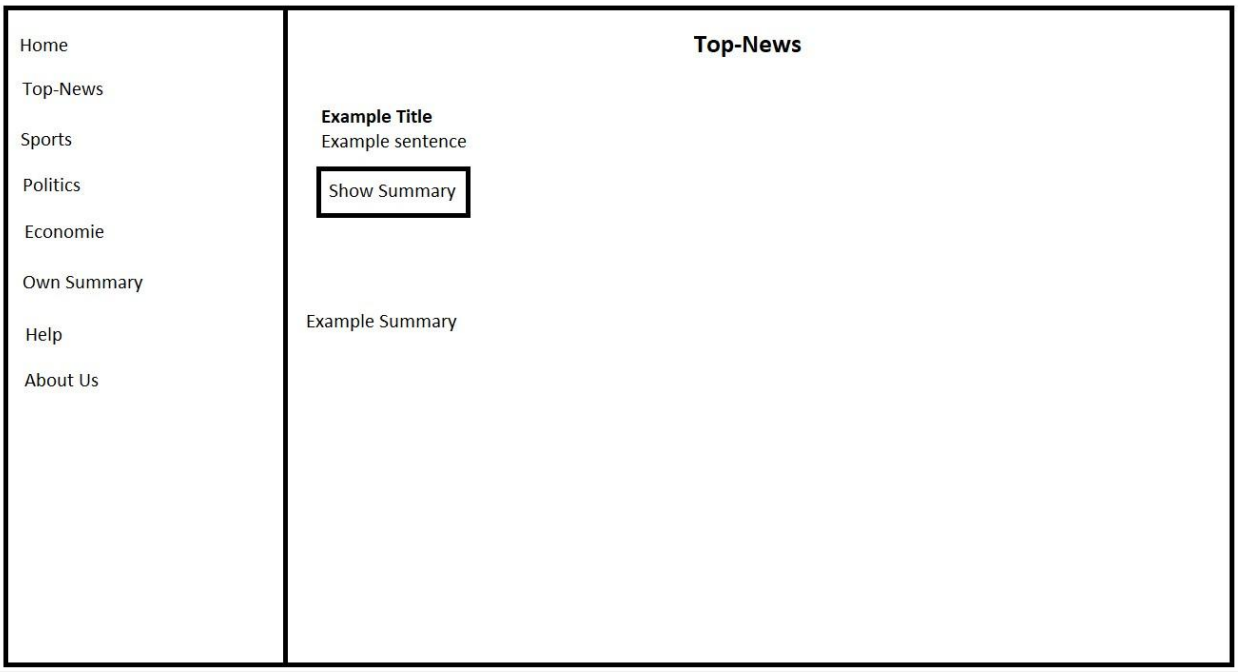
We are currently still waiting for feedback from the ML team regarding the details of the machine learning model. While we wait for feedback, we've searched and prepared datasets that could be used for training:

- https://huggingface.co/datasets/cnn_dailymail
- https://tagoyal.github.io/zeroshot-news-annotations.html

News Fetching

We will use web scraping with BeautifulSoup to obtain news articles from news sites.

1. First Prototype View



Same structure for different categories. For simplicity, only one example was created, because the general structure stays the same.

Home

Top-News

Sports

Politics

Economie

Own Summary

Help

About Us

Own Summary

Plese insert your article you want to have summarized:

Generate Summary