METHOD 1 (Basic Approach) - Matching Marketing Claims to Clinical Evidence using TF-IDF and Cosine Similarity

Rahul Manikandan

April 12, 2025

1. Objective

The goal of this project is to develop a fact-checking pipeline that matches a list of marketing claims with supporting evidence from clinical documents. The task involves analyzing documents related to the flu vaccine `Flublok` and identifying relevant textual excerpts that support each claim.

2. Methodology

2.1 TF-IDF and Cosine Similarity

The project uses the **TF-IDF (Term Frequency-Inverse Document Frequency)** vectorization technique combined with **cosine similarity** to compute the relevance of clinical document text with respect to each marketing claim.

- **TF-IDF** transforms text into numerical feature vectors based on word importance.
- **Cosine similarity** measures the angle between two vectors, giving a score between 0 and 1 (higher means more similar).

2.2 Pipeline Overview

The solution consists of the following steps:

1. Load marketing claims from a JSON file.
2. Load and extract text from clinical PDFs.
3. Clean the text (removing headers, URLs, newlines).
4. Build TF-IDF vectors from the clinical texts.

5. For each claim, compute cosine similarity against all clinical documents.
6. Select the top-K matches for each claim and export results in JSON format.

3. Source Code Structure

3.1 `preprocess.py`

Responsible for:

- Loading the marketing claims JSON.
- Listing and reading clinical PDF files.
- Extracting text from PDFs using `PyMuPDF`.
- Cleaning the extracted text using regex and whitespace normalization.

3.2 `matcher.py`

Handles the core logic of matching:

- Converts cleaned texts to TF-IDF vectors using `scikit-learn`.
- Calculates cosine similarity between each claim and all clinical texts.
- Returns the top-K most relevant document snippets per claim.

3.3 `utils.py`

Provides utility functions such as:

- Logger setup for consistent console output.
- JSON result writing function.

4. Project Structure

```
solstice-fact-check/
 data/
    Clinical Files/           # Clinical PDFs
    Flublok_Claims.json       # Marketing claims
 src/
    preprocess.py
    matcher.py
    utils.py
    __init__.py
```

```
results/
   basic_results.json
notebooks/
   main1.ipynb
requirements.txt
README.md
```

5. Requirements

To run this project, install dependencies with:

```
pip install -r requirements.txt
```

Main libraries used:

- `PyMuPDF` for PDF parsing.
- `scikit-learn` for TF-IDF and cosine similarity.
- `pandas, numpy` (optional for future extensions).

6. Results and Observations

The output is a JSON object where each marketing claim is matched with 3 clinical document snippets. Each snippet includes the document name, a short text excerpt, and a similarity score.

Why Scores Are Not High

TF-IDF only captures surface-level word frequency and does not understand semantics or paraphrasing. Clinical documents often use technical terminology or structure content across multiple pages, making it difficult for TF-IDF to detect strong overlaps with short, plain-language marketing claims.

- Scores mostly range between **0.1 to 0.3**.
- This is expected behavior and provides a good baseline for comparison.
- Future improvements can include semantic embeddings (e.g., `SBERT`, `OpenAI` embeddings).

```
{
    "claims": [
        {
            "claim": "Flublok ensures identical antigenic match with WHO- and FDA-selected flu strains.",
            "match_source": [
                {
                    "document_name": "FlublokPI.pdf",
                    "matching_text": "HIGHLIGHTS OF PRESCRIBING INFORMATION These highlights do not include all
                    "score": 0.12143134130269985
                },
                {
                    "document_name": "Treanor et al. (2011).pdf",
                    "matching_text": "Vaccine 29 7733\u2013 7739 Contents lists available at ScienceDirect Vacci
                    "score": 0.07467662456769332
                },
                {
                    "document_name": "Arunachalam et al. (2021).pdf",
                    "matching_text": "REVIEW ARTICLE OPEN Unique features of a recombinant haemagglutinin in\ufb
                    "score": 0.04184044073514859
                }
            ]
        },
```


Figure 1: Sample output showing top-3 matched clinical documents for selected marketing claims.

7. Sample Output Preview

8. Conclusion

This project successfully implements a basic NLP pipeline to match marketing claims with supporting clinical evidence. Despite limitations of TF-IDF in semantic understanding, the results demonstrate reasonable recall and provide a foundation for more advanced models in future work.