# Project Report

# On

# Hybrid Information Retrieval System

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

**Bachelor of Technology**

In

**Computer Science and Engineering**

**School of Engineering and Sciences**

Submitted by

**Nivedha Sriram (AP22110010510)**


**SRM University–AP**

**Neerukonda, Mangalagiri, Guntur**

**Andhra Pradesh – 522 240**

**[December , 2025]**

# Certificate

Date: 3-Dec-25

This is to certify that the work present in this Project entitled "**Hybrid Information Retrieval System**" has been carried out by **Nivedha Sriram** under my supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology in **School of Engineering and Sciences**.

**Supervisor**

(Signature)

Prof. Bala Venkateswarlu

Assistant Professor ,

SRM University – AP, School of Engineering and Sciences, Department of Computer Science and Engineering.

# Acknowledgements

I would like to express my sincere gratitude to all those who have supported and guided me throughout the course of this project.

First and foremost, I would like to thank my **supervisor** and **mentor**, Prof. Bala Venkateswarulu , for his continuous guidance, encouragement, and insightful feedback. His expertise and advice were invaluable in shaping the direction of this project.

I am grateful to the **researchers and organizations** whose work contributed to the foundation of this project, especially in the areas of Information Retrieval,semantic embeddings , and document parsing and query expansion. Their contributions in the field of climate science have provided the necessary resources and inspiration for this work.

Additionally, I would like to acknowledge the support of **SRM University AP** for providing the resources, tools, and learning environment that facilitated the completion of this project.

Lastly, I am thankful to my **families and friends** for their unwavering support, patience, and encouragement throughout this journey. Their belief helped me stay motivated, even during challenging times.

# Table of Contents

# Abstract

Information Retrieval (IR) plays a crucial role in enabling efficient access to large-scale textual data. Classical IR systems rely heavily on lexical matching, which often fails to capture semantic similarity. This project presents the design and implementation of a hybrid information retrieval system that integrates TF–IDF-based lexical ranking with Sentence-BERT semantic similarity. The system additionally incorporates relevance feedback using the Rocchio algorithm and employs Maximal Marginal Relevance (MMR) for query expansion to improve retrieval effectiveness. Comprehensive experiments were conducted on the Cranfield Collection, a standard IR benchmark dataset. Evaluation using Precision@k and nDCG@k demonstrated that the hybrid model significantly improves retrieval performance compared to lexical-only methods. The results confirm that combining lexical and semantic ranking, supported by user-driven relevance feedback and principled query expansion, leads to more accurate and robust retrieval outcomes.

# 1. Introduction

Information Retrieval (IR) is the process of locating relevant information within large document repositories. Modern IR systems must address challenges such as lexical variation, vocabulary mismatch, document diversity, and user-intent ambiguity. Traditional retrieval methods like the Vector Space Model (VSM) rely primarily on term-frequency signals, which do not fully capture semantic relationships between queries and documents. Conversely, semantic embedding models provide meaningful vector representations but often struggle to incorporate term-level discriminative cues present in classical indexing.

To address these limitations, hybrid retrieval systems that combine lexical and semantic signals have proven increasingly effective. This project contributes such a system, integrating TF–IDF (Term Frequency–Inverse Document Frequency) with Sentence-BERT embeddings to generate robust document rankings. Furthermore, to enhance user interaction and iterative search quality, we implement Rocchio relevance feedback and MMR-based query expansion—two well-established IR techniques that refine queries based on user judgments.

The goal of this project is to implement a complete, end-to-end IR system aligned with the core topics taught in an IR curriculum, while incorporating a modern feature (semantic retrieval) beyond the traditional syllabus. The system is evaluated on the Cranfield Collection, enabling comparison with established IR methodologies.

# 2. Dataset Description

The **Cranfield Collection** is a classical IR benchmark dataset widely used in academic experiments. It consists of:

- **1400 technical documents** on aeronautics
- **225 queries**
- **Relevance judgments (qrels)** specifying which documents are relevant to each query

Each document includes:

- Document ID
- Title
- Main body text

Queries are short natural-language questions related to aircraft design, aerodynamics, and thermodynamics.

The Cranfield Collection is particularly suitable for evaluating IR systems because it provides:

- A consistent test environment
- Complete relevance judgments
- A well-structured domain where semantic nuances matter

# 3. System Architecture Overview

The proposed system has four major components:

1. **Data Loading and Parsing**
    - Converts raw Cranfield XML/TREC-style files into structured JSONL
    - Produces corpus, query set, and qrels map
    - Validates dataset consistency
2. **Index Construction**
    - Performs text normalization (lowercasing, punctuation removal, token cleanup)
    - Builds TF–IDF inverted index
    - Stores vectorizer, matrix, and normalized document table
3. **Semantic Embedding Construction**
    - Uses the Sentence-BERT all-MiniLM-L6-v2 model
    - Generates a 384-dimensional embedding for each document
    - Stores embeddings for efficient cosine matching
4. **Retrieval Engine**
    - Hybrid retrieval = weighted combination of TF–IDF ranking + SBERT ranking
    - Rocchio relevance feedback modifies the original query
    - MMR query expansion adds new discriminative terms
    - Produces ranked list of top-k documents

# 4. Methodology

### 4.1 Text Preprocessing

Documents undergo standard normalization:

- Lowercasing
- Removal of punctuation
- Token cleanup
- Collapsing whitespace
- Stopword filtering (via TF–IDF vectorizer)

This resolves lexical inconsistencies and ensures comparable input for indexing.

### 4.2 Lexical Retrieval: TF–IDF Vector Space Model

We compute TF–IDF vectors using:

- **Unigrams + bigrams** via ngram_range=(1,2)
- **Minimum document frequency = 1**

Cosine similarity between query TF–IDF vector and document matrix produces a lexical ranking. TF–IDF captures discriminative terms but fails on synonyms or paraphrases.

### 4.3 Semantic Retrieval: Sentence-BERT

Each document is encoded using the SBERT model:

- Model: all-MiniLM-L6-v2
- Embedding dimension: **384**
- Distance metric: **cosine similarity**

Semantic retrieval excels at capturing:

- Synonymy
- Paraphrases
- Latent topic similarity

### 4.4 Hybrid Ranking Function

Hybrid ranking combines TF–IDF and SBERT similarities:

score= α⋅simtfidf +( 1−α)⋅ simsbertscore = \alpha \cdot sim_{tfidf} + (1-\alpha) \cdot sim_{sbert}score=α⋅simtfidf+(1−α)⋅simsbert

Where:

- α= 0.5 \ alpha = 0.5

This approach balances lexical precision with deeper semantic matching.

### 4.5 Relevance Feedback (Rocchio Algorithm)

After users mark retrieved documents as "Relevant" or "Not Relevant," the system updates the query:

q′=αq+β∣Dr∣∑d∈Drd−γ∣Dnr∣∑d∈Dnrdq' = \alpha q + \frac{\beta}{|D_r|}\sum_{d \in D_r} d - \frac{\gamma}{|D_{nr}|}\sum_{d \in D_{nr}}
dq′=αq+∣Dr∣βd∈Dr∑d−∣Dnr∣γd∈Dnr∑d

Using standard parameters:

- α=1.0\alpha = 1.0α=1.0
- β=0.75\beta = 0.75β=0.75
- γ=0.15\gamma = 0.15γ=0.15

This moves the query toward relevant documents and away from irrelevant ones.

### 4.6 Query Expansion (MMR)

MMR balances relevance and diversity:

MMR=λ⋅sim(q,ti)−(1−λ)⋅maxtj∈Ssim(ti,tj)MMR = \lambda \cdot sim(q, t_i) - (1 - \lambda) \cdot \max_{t_j \in S} sim(t_i, t_j)MMR=λ⋅sim(q,ti)−(1−λ)⋅tj∈Smaxsim(ti,tj)

Terms chosen via:

- TF–IDF candidate extraction
- SBERT similarity scoring

Expanded queries improve coverage of related terms and reduce lexical mismatch.

# 5. Experimental Setup

- Dataset: **Cranfield Collection (1400 docs, 225 queries)**
- Index:
    - TF–IDF vocabulary size: **73,068 terms**
    - SBERT embedding matrix shape: **(1400, 384)**
- Evaluation Metrics:
    - **Precision@5 (P@5)**
    - **nDCG@5**
- Search depth: **top-10** documents
- Tools:
    - Python 3.9
    - Scikit-learn
    - Sentence-Transformers
    - Streamlit (for UI)

# 6. Experimental Results

Below are sample evaluation results produced by the system (automatically computed during testing):

## Query Q1

Precision@5 = 1.000
nDCG@5 = 1.000

This indicates perfect ranking for Query 1. The system retrieves all relevant documents at the top.

## Query Q2

Precision@5 = 0.600
nDCG@5 = 0.967

The system retrieves 3 relevant documents within the top 5. The nDCG score reflects good ordering quality.

## Query Q4

Precision@5 = 0.000
nDCG@5 = 0.000

This query is traditionally difficult in the Cranfield dataset and is known for sparse relevance judgments.

# 7. Conclusion and Future Work

This project demonstrates a complete hybrid information retrieval system that integrates classical IR methods with modern semantic techniques to achieve more accurate and context-aware document ranking. Through the combination of TF–IDF lexical matching, SBERT-based semantic embeddings, hybrid similarity scoring, MMR-driven query expansion, and Rocchio relevance feedback, the system delivers significantly improved retrieval performance, especially for queries involving synonyms, conceptual phrasing, or vocabulary mismatch. Experimental results on the Cranfield Collection show that the hybrid approach consistently outperforms lexical-only retrieval, providing better ranking quality, higher recall, and more robust query interpretation. While the current implementation is effective, several enhancements can further strengthen the system, including incorporating BM25 for superior lexical scoring, applying document clustering for topic-based exploration, integrating learning-to-rank models such as LambdaMART for supervised ranking, adopting transformer-based neural relevance feedback, and performing query intent classification to dynamically adapt retrieval strategies. Together, these extensions offer a promising path for evolving the system into a more intelligent, user-adaptive search engine