**Project Report-**

**Title:**

Fact-Checking and Information Verification System

**Objective:**

The goal of this project is to build a system that can:

1. Extract named entities from a given text.
2. Compare a claim against a database of fact-checked claims using **TF-IDF** and **cosine similarity**.
3. Compute semantic similarity between sentences using **SentenceTransformer**.
4. Query external APIs (e.g., Wikidata) for additional information.
5. Load and analyze a dataset for question-answering tasks.

**Components of the Project:**

1. **Named Entity Recognition (NER)**:
   * Using spaCy, the system extracts named entities (e.g., people, organizations, dates) from a given text.
   * Example:

Named Entities: [('Elon Musk', 'PERSON'), ('Tesla', 'ORG'), ('2025', 'DATE')]

1. **TF-IDF and Cosine Similarity**:
   * The system compares a claim against a database of fact-checked claims using **TF-IDF** and **cosine similarity**.
   * Example-

Cosine Similarity (TF-IDF): [[0.70710678 0.70710678]]

1. **Semantic Similarity**:
   * Using the **SentenceTransformer** model (all-MiniLM-L6-v2), the system computes semantic similarity between sentences.
   * Example:

Semantic Similarity (SentenceTransformer): tensor([[0.8765]])

1. **Wikidata API Query**:
   * The system queries the **Wikidata API** to retrieve information related to a specific query.
   * Example:

Copy

Wikidata API Response: {'searchinfo': {'search': 'Tesla Model release 2025'}, ...}

1. **Dataset Analysis**:
   * The system loads the **SQuAD dataset** and analyzes its structure.
   * Example:

Column names in the dataset:

['id', 'title', 'context', 'question', 'answers']

**Dataset Used:**

* **SQuAD (Stanford Question Answering Dataset)**:
  + A reading comprehension dataset consisting of questions posed by crowdworkers on a set of Wikipedia articles.
  + Contains 100,000+ question-answer pairs.
  + Available on Hugging Face's datasets library.
  + Source: [SQuAD Dataset on Hugging Face](https://huggingface.co/datasets/squad" \t "_blank)

**Results:**

1. **Named Entities**:
   * Successfully extracted entities such as Elon Musk (PERSON), Tesla (ORG), and 2025 (DATE).
2. **Cosine Similarity**:
   * The extracted claim showed high similarity (0.707) with the fact-checked claims.
3. **Semantic Similarity**:
   * The semantic similarity between the two sentences was 0.8765, indicating high similarity.
4. **Wikidata API**:
   * Retrieved relevant information about the query "Tesla Model release 2025".
5. **Dataset Analysis**:
   * The SQuAD dataset contains columns: id, title, context, question, and answers.

**Conclusion:**

This project demonstrates a pipeline for fact-checking and information verification using NLP techniques. It combines:

* Named entity recognition (NER) for extracting key information.
* Text similarity (TF-IDF and semantic similarity) for comparing claims.
* External APIs (Wikidata) for retrieving additional information.
* Dataset analysis for understanding the structure of question-answering datasets.

The system can be extended to handle more complex claims and integrate with larger fact-checking databases.

**Difficulties Faced While Making This Project**

While building this project, several challenges may arise,

* **Difficulty**: Installing and configuring libraries like spaCy, scikit-learn, sentence-transformers, and datasets can be tricky, especially if there are version conflicts or missing dependencies.
* Difficulty: struggle to understand concepts like Named Entity Recognition (NER), TF-IDF, cosine similarity, and semantic similarity.
* Difficulty: Querying external APIs (e.g., Wikidata) may fail due to incorrect URLs, or API rate limits.
* Running NLP models (e.g., spaCy or SentenceTransformer) on large texts or datasets can be slow.
* struggle to find appropriate datasets or understand their licensing terms

**Learning -**

As a college student, I often found myself limited to theoretical knowledge without understanding how to apply it in real-world scenarios. This internship program completely transformed my perspective by starting from the basics and fundamentals, and guiding me through the entire process of data collection, visualization, and efficient utilization using machine learning algorithms. It not only enhanced my technical skills but also built my confidence to lead teams and collaborate effectively.

**Key Learnings from the Program:**

1. **Bridging Theory and Practice**:
   * Learned how to apply theoretical concepts to real-world problems.
   * Understood the end-to-end process of data science, from data collection to actionable insights.
2. **Technical Skills Development**:
   * Gained hands-on experience in data collection, cleaning, and visualization.
   * Applied machine learning algorithms to solve practical problems.
   * Worked on tools and technologies used in the industry.
3. **Leadership and Teamwork**:
   * Developed leadership qualities by leading team projects and collaborating with peers.
   * Improved my ability to communicate ideas effectively and work in a team-oriented environment.
4. **Practical Project Experience**:
   * **Music Recommendation System using Clustering**: Led a machine learning project to build a recommendation system, applying clustering techniques to analyze user preferences.
   * **Skill Clustering and Automated Code Scanning System**: Worked on projects that involved clustering skills
   * automating code analysis for efficiency.
   * **Competitor and Financial Analysis**: Performed in-depth competitor and financial analysis for a company, using advanced data analysis techniques to deliver impactful insights.
   * **Real-Time Fact-Checking System**: Contributed to a real-world project focused on building a real-time fact-checking system, which involved processing and analyzing large datasets.
5. **Confidence Building**:
   * The program helped me gain confidence in my technical abilities and leadership skills.
   * Working on real-world projects gave me a sense of accomplishment and prepared me for future challenges.

**How This Program Helped Me:**

This internship provided me with the opportunity to work on practical, industry-relevant projects, which is often missing in traditional college education. It taught me how to approach problems systematically, use data effectively, and deliver results that have a real-world impact. The experience not only enhanced my technical expertise but also honed my soft skills, such as teamwork, leadership, and communication, making me a well-rounded professional.