Project Summary: AI/ML Solution for Research Gathering in Continuous Manufacturing

The project focuses on developing an AI/ML-powered system to automate the gathering, parsing, and categorization of research related to continuous manufacturing. Instead of using web crawlers, the system utilizes API keys to fetch metadata from 16 different data sources, including journals, books, documents, and workshops. The goal is to create a system that continuously collects and categorizes the latest research, providing industry professionals with up-to-date and relevant information.

System Workflow:

1. **Step 1: Data Extraction**
    - Metadata is extracted from 16 trusted data resources through APIs, focusing on records from the current month.
    - Primary filtering is applied, where only records with titles, abstracts, and keywords containing predefined core keywords are included.
2. **Step 2: Keyword Normalization**
    - Core tags (e.g., "API," "Oral Solid Dosage") are defined, and terms with similar meanings (e.g., "Active Pharmaceutical Ingredient" → "API") are grouped.
    - Semantic clustering and normalization ensure consistency and relevance, allowing the system to categorize data more effectively.
3. **Step 3: Preprocessing for Area Prediction**
    - Dummy variables are created for each area label (e.g., "API," "Oral Solid Dosage").
    - Embedding models (e.g., BioBERT, SBERT) are trained to predict area labels for research documents, which are then used as features for machine learning models.
4. **Step 4: Area Prediction**
    - Machine learning models, such as Random Forest and Logistic Regression, are used to predict area labels based on embeddings.

- Separate models are trained for each label, with techniques like SMOTE applied to address label imbalance and optimize model performance.

Methodology:

The project is divided into several key phases:

1. **Data Collection**: Metadata is gathered using API keys, fetching data from 16 different sources such as journals, conference proceedings, and patents, with a focus on the latest research related to continuous manufacturing.
2. **Embedding Model Fine-Tuning**: Embedding models are fine-tuned for six specific columns of research data. These models are trained to convert textual data into vector representations, which are used in machine learning models for categorization and classification.
3. **Model Training**: Machine learning models are trained using the embeddings generated in Step 2. These models are designed to automatically categorize and classify research data into predefined taxonomies, making it easier for users to access relevant information.
4. **Regulatory Focus**: For the **Regulatory** column, a specialized approach is taken due to the limited data volume. A **BioBERT-based model** is fine-tuned to assess the relevance of documents, improving classification accuracy for this specific category.

Results:

The AI/ML solution has successfully automated the research-gathering process, enabling continuous collection, parsing, and categorization of research data. The methodology applied ensures that the system is able to process large volumes of data efficiently while maintaining high accuracy in categorization. The embedding models have been fine-tuned to classify research into key areas such as "API," "Oral Solid Dosage," and "Regulatory," with the **BioBERT model** particularly improving the classification of regulatory documents. Additionally, the system performs well in addressing label imbalances through oversampling techniques like SMOTE, resulting in a more balanced and accurate model.

The AI models now classify documents based on relevance, using embeddings to represent complex relationships in research data. The system is designed to continuously learn, making it adaptable to new research trends and emerging topics in continuous manufacturing.

Impact on Improving Research Gathering:

1. **Increased Efficiency**: The automation of data extraction and categorization significantly reduces the time and effort required to manually gather and organize research. Industry professionals can access the latest findings quickly, allowing them to stay ahead of trends and innovations in continuous manufacturing.
2. **Improved Categorization**: By leveraging AI/ML models, the system ensures accurate and consistent categorization of research data. This leads to a more organized and searchable database, improving the usability of research for end-users.
3. **Dynamic and Scalable**: The system's ability to continuously update and adapt to new research ensures it remains relevant as the field of continuous manufacturing evolves. It can scale with the growth of research in this field and dynamically adjust to new topics and categories.
4. **Focused on Regulatory Research**: By applying specialized models like **BioBERT**, the project has addressed the challenge of limited data in regulatory research. This ensures more accurate classification of documents in this critical area, which is often a challenge for traditional machine learning methods.

Final Deliverables:

The final system will include a user-friendly AI/ML solution for continuously gathering and categorizing research in continuous manufacturing. Key deliverables include:

- A comprehensive, dynamic, and scalable AI-powered system that categorizes research.
- Detailed system documentation and a recorded demo showing how the system works.
- A summary report detailing the methodology, results, and the impact of the system on improving research gathering.

This solution will provide a valuable resource for professionals in the continuous manufacturing industry, helping them stay informed about the latest research and advancements while improving the efficiency and accuracy of their research-gathering processes.