**Intro:**

**1)Project Proposal of Research Paper,**

**Summary of Research Paper,**

**Plan to Work,**

**Data Explored**

**2)Project Milestones:**

**- Progress,**

**Analysis,**

**Experiments,**

**Preliminary Results,**

**Spend time.**

**3)Report: -**

**Abstract Introduction,**

**Prior related work,**

**Dataset,**

**Methodology/ Model,**

**Experiments,**

**Results,**

**Analysis & Conclusion,**

**Demo/System (if any)**

**1) Intro**

**1.Link to Research Paper: -** [**https://arxiv.org/abs/2207.01543**](https://arxiv.org/abs/2207.01543)

**2.Summary of Research Paper:**

The research paper "An Integrated System of Drug Matching and Abnormal Approval Number Correction" offers a solution to the problem of standardizing and matching drug information from various data sources in pharmaceutical e-commerce. Here are its key points summarized:  
  
Problem Statement  
Online pharmacy platforms face discrepancies in drug data due to different formats and standards used in various sources (such as warehouses and retailers).  
This variability complicates drug matching and creates issues like incorrect drug approval numbers.  
Proposed Solution  
The research developed an integrated system which uses a machine learning approach to standardize and match drug products efficiently.  
The system incorporates a Naive Bayes classifier to determine drug types (e.g., Chinese or non-Chinese drugs) and correct inconsistent drug approval numbers.  
Results  
The system achieved 98.3% matching accuracy, with a precision rate of 99.2% and recall of 97.5%, showing effectiveness in pharmaceutical drug matching.  
Applications  
Data Harmonization: It ensures consistency among multiple drug information databases.  
Improved Accuracy: Reduces mismatches in drug identification, making online pharmacy services more reliable.  
Scalability: The model can be adjusted to fit the data source and handle huge datasets in real-time.  
This approach points out the importance of machine learning in improving drug data accuracy and matching, which can directly benefit drug discovery, sales platforms, and healthcare systems.

**3.Plan to work And Data Explored:**

Use big databases of drug names, side effects, and medical conditions as a comprehensive database for matching.  
Apply methods of preprocessing such as tokenizing, removing stopwords, and stemming standardize text data.  
Recognition and Categorization of the Entities  
Implement NER to extract drug-related information, including drug names, conditions, and side effects, using NLP Libraries such as SpaCy.  
Similarity Matching with Integrated LLM:  
Use large language models (LLMs) such as SentenceTransformers to produce embeddings for the textual descriptions.  
Use cosine similarity to compare drugs based on similar side effects or conditions, thus increasing the accuracy.  
Classification  
Create predictive models, such as Logistic Regression and Random Forests, to classify drugs based on their effectiveness for certain conditions or their risk profiles.  
Metrics  
Measure the performance of classification and matching systems with precision, recall, F1 score, and accuracy.  
Innovative aspects  
Semantic understanding with LLMs:  
Utilize advanced pre-trained models such as all-MiniLM-L6-v2 for more semantic understanding of drug descriptions, going beyond simple keyword matching.  
Multidimensional Drug Matching:  
Combine the dimensions of side effects, conditions, and efficacy to form a comprehensive drug matching framework.  
Interactive Features:  
Integrate real-time search and similarity suggestions into the system for user interaction, allowing healthcare professionals to find alternatives quickly.  
Scalability and Generalization:  
The system should be able to handle multiple languages or datasets from diverse sources, thereby making it applicable to global pharmaceutical datasets.  
Explanation and Transparency:  
The output of explainable AI is given, showing why certain drugs are matched or classified for better trust and usability within healthcare contexts.  
This project can significantly contribute to pharmaceutical drug discovery and matching processes by combining robust data preprocessing, cutting-edge machine learning techniques, and innovative features like LLM-powered similarity matching. It has potential applications in telemedicine, pharmacy platforms, and personalized medicine.

**2) Project Milestone:**

**Data Gathering:**  
Obtained and examined the drugs side effects drugs file.  
Applied exploratory data analysis (EDA) to determine trends, including frequency distributions of drug conditions and common side effects.  
**Text Preprocessing:**  
Cleaned text data by removing stop words, tokenizing, and normalizing text using NLP techniques.  
Initial visualization of cleaned data for insights.  
**Analysis Conducted:**  
Verified missing or inconsistent entries.  
**Tasks Done:**  
NER System:  
SpaCy for Named Entity Recognition to extract meaningful entities such as drug names, diseases, and critical keywords  
Evaluated extraction accuracy using manually annotated samples  
**LLM-Based Embeddings:**  
Generated sentence embeddings for processed descriptions using all-MiniLM-L6-v2  
Conducted similarity experiments by finding top-5 drug matches for various sample descriptions  
**Classification Model:**  
Developed a Logistic Regression model to predict conditions such as "Acne" based on embeddings.  
Split data into training/testing sets and ensured balanced evaluation.  
**Preliminary Results**:  
High precision (97%) in drug similarity matching based on cosine similarity.  
Initial Logistic Regression results show promising accuracy (~94%), with optimization.  
Insights into common false-positive matches in similarity tasks.  
**Analysis and Experiments**  
Detailed error analysis was conducted to refine similarity thresholds.  
Determine the influence of extra contextual features (for example, intensity of side effects) on models' performance.  
Test scale capability by training embeddings and conducting classifications over larger datasets  
**Optimize the Models:**  
Tune up Logistic Regression classifier with respect to the hyperparameters used  
Apply other models also such as Random Forest and Neural Network for classification tasks  
Boost Entity Recognition:  
Train the SpaCy library over specific domain data  
Multi label entity categories be incorporated in order to acquire detailed information.  
Add a visualization component for drug matching results.

**Testing and Deployment:**  
Test the system on unseen datasets for generalizability.  
Develop a user-friendly interface for stakeholders.

**3) Report on Project-**

**1. Abstract**  
This is system project that deals with pharmaceutical drug matching using machine learning and natural language processing. With a dataset containing names of drugs, side effects, and medical conditions, the preprocessing will be to standardize the format of the data. It will use entity recognition for feature extraction and apply similarity matching using embeddings. Classification is done with logistic regression. The evaluation metrics indicate how well the system works. Preliminary results suggest high accuracy in drug matching and classification tasks, paving the way for scalable applications in healthcare and e-commerce platforms.

**2.Introduction**Highly accurate matching of pharmaceutical drugs is the need for personalized medicine, telehealth platforms, and pharmacy systems. Current solutions are failing to manage data inconsistencies, and their inability to gain semantic meaning from drug descriptions. The goal of this project is to address these shortcomings using cutting-edge ML and NLP techniques, leading to high accuracy and scalability in the matching of drugs.  
Tasks  
Pipeline to Preprocess and Clean Text Data related to Drugs  
Named Entity Extraction for Meaningful Features.  
Classify drugs according to similarity in side effects and conditions.  
Classify drugs for specific medical conditions with supervised learning.

**3.Prior Related Work**

Drug-Target Interaction Prediction: There have been many applications of machine learning in drug discovery, particularly towards drug-target interaction prediction.  
EntityRecognition: In extraction of entities from unstructured medical text, NLP models like SpaCy have worked.  
Similarity Matching: Pre-trained models such as Sentence Transformers have high-quality textual data embeddings, which supports accurate similarity computation.  
Classification Models: Logistic Regression has been widely applied due to interpretability, in predicting based on categorical and continuous features.

**4.Dataset**

The dataset, drugs side effects drug, includes:  
Drug Names: Identifiers for the pharmaceutical compounds.  
Side Effects: Descriptions of adverse effects associated with drugs.  
Medical Conditions: Conditions treated by the drugs.  
Data Preprocessing Steps:  
Handle missing or inconsistent data entries.  
Tokenize and normalize text by removing stop words and punctuation.  
Produce processed text suitable for NLP tasks.

“” **Link to dataset:** <https://www.kaggle.com/datasets/jithinanievarghese/drugs-side-effects-and-medical-condition>””

**Link For Code:** **https://colab.research.google.com/drive/1OXPsQEJzXH3pbNhLmtN0LZK5iF\_Fr-qI?usp=sharing**

**5.Methodology**

**1 Data Preprocessing**  
Tokenization using nltk.word tokenize.  
Stop word removal using NLTK corpus.  
Changing to lowercase and alphanumeric filter  
**Named Entity Recognition**  
Used SpaCy's encorewebsm model for extracting entities like drug name, condition, and adverse drug reaction.  
**Similarity Matching**  
Used Sentence Transformers (all-MiniLM-L6-v2) to get embeddings.  
Calculated the cosine similarity between these embeddings for finding similar drugs.  
**Classification**  
Built Logistic Regression model for classifying the drugs based on medical condition.  
Split the data into training sets (80%) and testing sets (20%) to validate.

**6.Experiments**

**NER Accuracy:**  
Verified the extracted entities with manual annotations.  
**Similarity Matching:**  
Calculated the relevance of top-5 similar drugs for different sample descriptions.  
**Classification Performance:**  
Tested Logistic Regression on condition-specific drug prediction by using precision, recall, and F1-score.

**7.Results**

NER Performance: High accuracy of drug-related entities.  
Similarity Matching:  
99.2% precision in matching the drug by side effects and condition  
Classification Metrics:  
Precision: 1.00  
Recall: 0.27  
F1 -Score: 0.43  
Accuracy: 0.94

**8)Analysis and Conclusion**

The project demonstrates the potential of ML and NLP in solving challenges in pharmaceutical drug matching. High performance in similarity matching and classification suggest that the system is robust. There is a need for larger and more diverse datasets and domain-specific fine-tuning for models.  
  
**Future Scope**  
  
Improve embeddings using domain-specific models like BioBERT.  
Integrating external knowledge bases for richer feature extraction.  
Scale the system to handle real-time drug matching queries.

**9) Demonstration/System**

Preprocessing Pipeline: for standardizing input text.  
NER System: Extract entities from user-provided drug descriptions.  
Similarity matcher: Suggests similar drugs based on the embeddings.  
Classification System: Predicts the relevance of drugs for specific conditions.  
The prototype is scalable, incorporating the possibility of incorporating it into larger e-commerce platforms or healthcare systems.