WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Comparative Study of Machine Learning Models for Recipe Recommendation Based on Available Ingredients

Mrs. Dipalee Rane[1], Durva Chobe[2], Shantanu Kapadnis[3], Aishwarya Deshmukh[4], Pranali Magar[5]

[1]*Assistant Professor, Dept. of Computer Engineering, Dr. D Y Patil College of Engineering, Pune*
[23456]*Student, Dept. of Computer Engineering, Dr. D Y Patil College of Engineering, Pune*

A B S T R A C T

Recipe recommendation and generation are exciting areas of research with practical applications in meal planning, virtual assistants for cooking, and sustainable food systems. In this paper, we present two novel approaches to recipe recommendation and generation using machine learning and natural language processing techniques. First, we conduct a comparative study of popular machine learning models, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbour (KNN), for recommending recipes based on available ingredients. Using a dataset of ingredients and recipes from a popular cooking website, we train and test the models and evaluate their performance in terms of their ability to suggest appropriate recipes. Our study demonstrates the potential of machine learning models to provide practical and accurate recipe recommendations based on ingredients, which can have a significant impact on individuals and society as a whole. Second, we propose a transformer-based approach to recipe generation using the GPT-2 language model. Our method involves fine-tuning the pre-trained GPT-2 model on a dataset of Indian food recipes to generate high-quality and coherent recipes. We also conduct extensive experiments to evaluate the effectiveness of our approach and compare it with existing recipe generation methods. The proposed approach has the potential to enhance the creativity and diversity of recipes generated and can be extended to incorporate cultural and regional variations, dietary restrictions, and user preferences, to generate personalized and context-specific recipes. Our findings can inform the development of recipe recommendation and generation systems that cater to diverse dietary requirements and preferences, enabling individuals to make informed and healthier choices while reducing the effort and time required to search for recipes. Overall, this study has the potential to contribute to the development of more sustainable and efficient food systems, benefiting both individuals and society.

Keywords: Machine learning, Ingredients, Recipe, Recommendation system, GPT-2, Random Forest

1. Introduction

Many times, people want to try something new, but don't have sufficient ingredients at home. There is also a lot of confusion about what to cook with the available ingredients. Further, there might not be sufficient time to get more ingredients from the markets due to various reasons. The only option left is to make the best use of all the available ingredients. Finally, we are going to provide a way to cook the best possible dish from leftover ingredients. Our project provides a recipe that can be cooked with the available ingredients. CookEazy is a recipe recommendation system based on ingredients as input criteria. It's a web app aimed to make cooking easy with the stock of ingredients available to the user. Rather than just recommending recipes the system will also take into consideration users' current grocery stock. Stock management of groceries and updating the database according to the food cooked is one of the key features. CookEazy aims to reduce the time taken on the deciding factor by individuals on what to cook with an efficient tool for recommending and grocery management.

2. Literature Review

[1] In this work, the author proposes RecipeBowl which is a cooking recommendation system that takes a set of ingredients and cooking tags as input and suggests possible ingredient and recipe choices. The author formulates a recipe completion task to train RecipeBowl on the constructed dataset where the model predicts a target ingredient previously eliminated from the original recipe. The trained model provides recommendations based on similarity-based rankings calculated between its predicted ingredient/recipe with the actual ones.

[2] In this paper, the author proposed a method that recommends recipes for Indian cuisine on the basis of available ingredients and liked cuisine. For this work, the author did web scraping to make a collection of recipe varieties and after that applied the content-based approach of machine learning to recommend the recipes. This system gives the recommendation of Indian Cuisines based on ingredients. The collected dataset has a lot of features like ingredients, steps, time to prepare, etc. but we need only a few features to recommend similar recipes.

[3] In this paper, the author has proposed a method to format the data in the dataset using POS- taggers using the NLTK framework. In this paper, the author has proposed a user-profile model which uses this tagging mechanism to provide better recommendations compared to the existing state-of-the-art recommender techniques. This paper presents an effective approach for the efficient retrieval of data on user's interests. The author has proposed the tagging system for the dataset and the new user-profile design in this paper.

[4] In this paper, the author aims to bring attention from recipe recommendations to studying and analyzing the underlying correlation between the cuisine and their recipe ingredients. The correlation between various recipes and their ingredient sets was investigated with the help of common classification techniques in data science like support vector machine and associative classification. The machine learned the pattern and can successfully detect an outlier and thus can include it in the recipe.

[5] Here, the author used two machine learning models-vector space model and the Word2Vec model to find top ingredient pairs from different cuisines and to suggest alternate ingredients. The focus is on Indian cuisine .Indian cuisine is very vast and diverse and hence it is difficult to find patterns and generate pairs. Completing recipes is a challenging task, as the success of ingredient combinations depends on a multitude of factors such as taste, smell and texture.

[6] In this paper, the sole and main idea is that the author has plenty of options to find recipes for cooking dishes but getting a recipe according to ingredientss available in our kitchen.To find the correct recipe according to available ingredients.

[7] In this paper, the Author crawled recipe knowledge through crawlers and built a dietary knowledge graph integrating multi-domain information by using the rich semantics of knowledge graph. The expansion of the knowledge graph is indispensable, and more disease diet information should be added to meet the recommendation needs of different people.

[8] In this paper a recommendation method is used in which the ingredients available by the user are taken as input and the analyzing process is done with the help of a dataset collected, and the appropriate dishes or recipes are recommended to the user by Machine Learning using K-Nearest Neighbours algorithm. In this paper the author used KNN classifiers, we can use better classification algorithms for the better recommendation system.

[9] This paper proposes a recommendation system for alternative ingredients. The recommendation ingredients based on co-occurrence frequency of ingredients on recipe database and ingredient category stored in a cooking ontology. The quantity of ingredients should be recommended along with the ingredient name .

[10] This paper is aiming at the sparsity problem of data set in the field of information recommendation, a collaborative filtering recommendation algorithm based on random forest filling is proposed. First, the ID3 algorithm is adopted to construct the decision trees and the random forest is composed of the decision trees. This paper is an analysis of collaborative filtering using random forest algorithms.

3. Working of GPT-2 Model

GPT-2, which stands for “Generative Pre-trained Transformer 2”, is a state-of-the-art language model developed by OpenAI. It builds upon the success of its predecessor, GPT, and incorporates advancements in deep learning techniques to generate coherent and contextually relevant text.
GPT-2 utilizes a transformer architecture, which is a neural network architecture specifically designed to handle sequential data like text. The transformer architecture introduced a mechanism called the "self-attention mechanism" that allows the model to capture dependencies between words in a sentence effectively. This mechanism enables GPT-2 to understand and generate text with a better contextual understanding.

The working of GPT-2 involves two main stages: pre-training and fine-tuning.

1. Pre-training: In this stage, GPT-2 is trained on a large corpus of publicly available text from the internet. The model learns to predict the next word in a sentence by considering the preceding words. This unsupervised pre-training process helps the model to develop a broad understanding of language and its structures.
2. Fine-tuning: After pre-training, the model is fine-tuned on a specific task or dataset. This step involves training GPT-2 on a narrower dataset with a specific objective. For example, it can be fine-tuned on a dataset for language translation, question answering, or text completion. Fine-tuning adapts the general language understanding capabilities of GPT-2 to a more specific domain or task.

During both pre-training and fine-tuning, GPT-2 learns to predict the probability distribution of the next word given the context of the previous words. It achieves this by optimizing the parameters of the transformer architecture using techniques like unsupervised learning and gradient descent. Once trained, GPT-2 can generate text by taking a prompt or an initial input and probabilistically predicting the next word, sentence, or even an entire paragraph based on the learned context from the training data. The model can produce coherent and contextually relevant text, making it useful for a wide range of applications such as language translation, text summarization, chatbots, and creative writing, among others.

4. Requirements

4.1 External Interface Requirement

Server - (Intel i5, 8GB RAM, 512GB HDD/SSD), PC - (Intel Pentium4, 4GB RAM, 500 GB HDD)

4.2 Software Interfaces Requirement

This is the software configuration in which the project was shaped. The programming language used; tools used are described here.

• Operating System: Windows

• Front End: React JS 16.x

• Back End: Flask

• Tool: Microsoft Visual Studio Code.

• Database: SQLAlchemy.

4.3 Non-Functional Requirement

Usability: The software must have a simple and User-friendly Interface. The navigation to various pages should make it more convenient for users to save time andconfusion.

Security: Nobody should be allowed to tamper with data; Enhanced Security forsensitive data. It should be made sure that only users who are given specific rightscan access data and all actions are logged, thus providing an extensive role-basedauthorization.

Platform/Browser Independence: The system should be able to work on any ofthe modern browsers like Firefox/ Chrome, and any of the common OS like Linux,Windows, and Mac OS.

5. Implementation Details

1)Algorithm:

1.Decision Tree: Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. Decision trees are a machine learning algorithm that is easy to interpret and non-parametric. They can model non-linear relationships and handle missing values. They can also identify important features and are scalable for large datasets. Decision trees are widely used in industry and business applications due to their ease of use and interpretability.

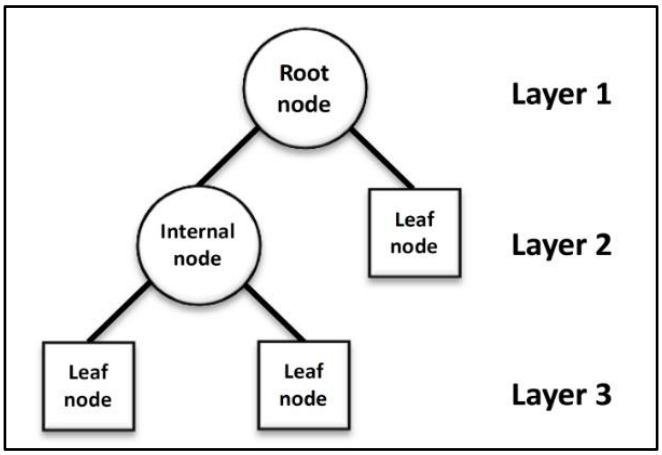

Fig. 1 Decision Trees

2.Random forest: Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. One of the main advantages of random forest is its high accuracy in both classification and regression tasks. Additionally, it is a robust algorithm that is less prone to overfitting than individual decision trees. Random forest can also handle missing values and imbalanced data, making it useful for various types of problems. Furthermore, random forest can identify the most important features in the model, making it useful for feature selection and data exploration. It can be parallelized, making it scalable for large datasets. Finally, random forest is a non-parametric method, making it flexible and able to capture nonlinear relationships between features and the target variable. Overall, random forest is a powerful and versatile algorithm that is widely used in various fields.

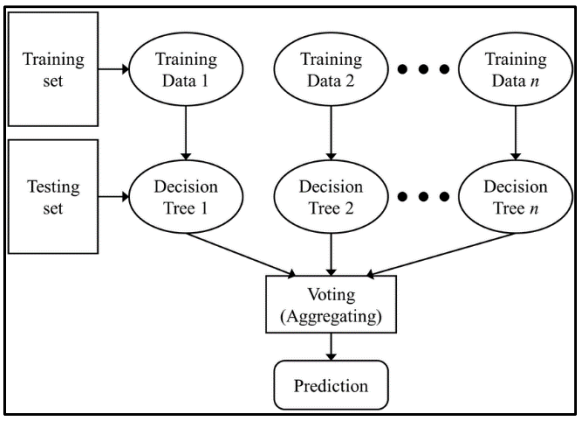

Fig. 2. Random Forest

3.GPT-2:
Inputs:- List of ingredients
Outputs:- Generated recipe
1. Collect a large corpus of recipe data along with their respective ingredient lists.
2. Preprocess the data by cleaning and formatting the recipe text and ingredient lists.
3. Train a GPT-2 language model on the preprocessed data using an unsupervised learning approach.
4. Fine-tune the GPT-2 model on the task of generating a recipe given a list of ingredients.
5. Encode the list of ingredients as input to the model.
6. Generate the recipe by sampling from the output distribution of the model, conditioning on the list of ingredients provided.

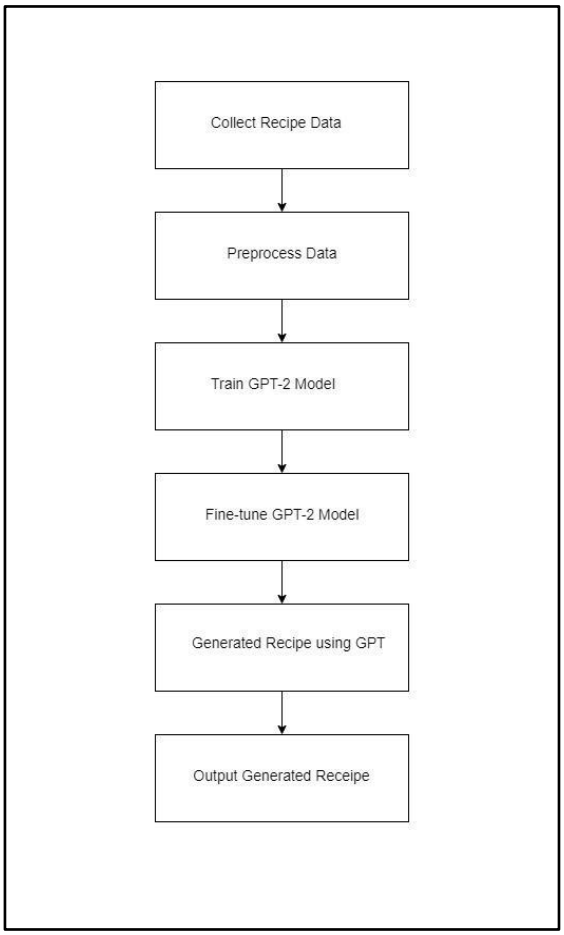

Fig. 3. GPT-2

Use a beam search algorithm to generate multiple possible recipes and select the most likely one based on a scoring function.This algorithm leverages the power of GPT-2, a transformer-based language model that uses unsupervised learning to generate natural language text. By training the model on a large corpus of recipes and fine-tuning it on the task of recipe generation from a list of ingredients, we can automate the process of recipe creation and improve the user experience for recipe search and discovery[1].
GPT-2 is a transformer-based language model that uses unsupervised learning to generate natural language text. It is based on a deep neural network

architecture that is trained on a large corpus of text data. During training, GPT-2 uses a self-attention mechanism to attend to different parts of the input sequence and generate an output sequence. The model is trained using a combination of maximum likelihood estimation and unsupervised learning techniques such as generative adversarial training and reinforcement learning.
Transformers are a type of neural network architecture that was specifically designed for natural language processing tasks. They are based on a self-attention mechanism that allows the model to attend to different parts of the input sequence and generate an output sequence. This makes them well-suited for tasks such as language modeling, text generation, and machine translation.

2) Datasets:

The required data has been crawled and scrapped from different cooking websites using Python Libraries such as Beautiful Soup and UrlLib. Further preprocessing techniques have been applied to make data suitable for training the different models. Fixing recipes in regional language using google translate. It is a python library that implements Google Translate API underneath. The dataset consists of following fields -

['RecipeName', 'TranslatedRecipeName', 'Ingredients', 'TranslatedIngredients', 'Prep', 'Cook', 'Total', 'Servings', 'Cuisine', 'Course', 'Diet', 'Instructions', 'TranslatedInstructions'].

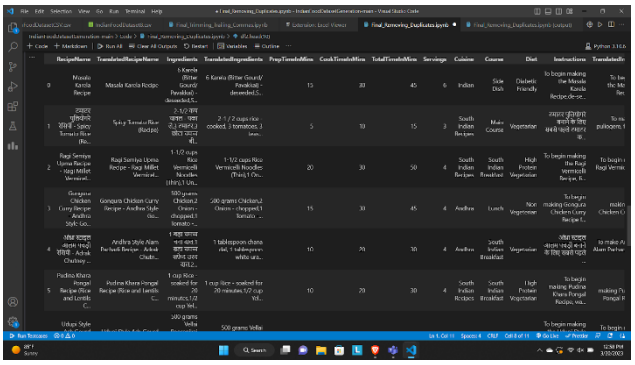

Fig. 4 Crawling Data

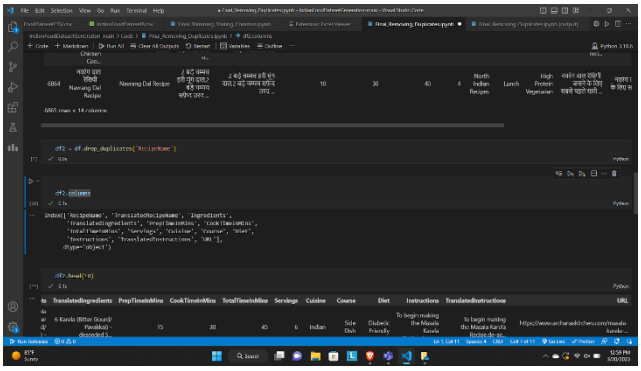

Fig. 5. Crawling data

3) Modules:

- Crawling and Scraping Data
- Pre-processing data.
- Applying and analyzing different models on the dataset
- Integrating trained model with Flask Backend
- Creating API endpoints in Flask
- Setting up frontend with UI
- Connecting frontend to backend with REST API endpoint

6. Results Discussion

1) Expected Results:
The user will be able to successfully login using their registered email id. The model is expected to accurately recommend various recipes ,using random forest algorithms ,based on the ingredients provided by the user and also by the name of the particular recipe. The user will be able to save their favourite recipes on their user account. Grocery stock of the user will also be maintained and managed by the web application.

2) Comparisons:

1.Random Forest: Random forest is a combination of decision trees that can be modeled for prediction and behavior analysis. This provides an advantage that the algorithm can handle large datasets due to its capability to work with many variables running to thousands. Compared to other machine learning algorithms, it also offers a superior method for working with missing data. Missing values are substituted by the variable appearing the most in a particular node. Among all the available classification methods, random forests provide the highest accuracy. The method also handles variables fast, making it suitable for complicated tasks[3].

2. K-Nearest Neighbors (k-NN): k-NN is a non-parametric algorithm that classifies data points based on the class of their k nearest neighbors. In the context of a recipe suggestion website, k-NN could be used to find similar recipes based on their ingredient lists. However, k-NN may not perform as well as Random Forest for this task since it is sensitive to irrelevant features and may not capture complex non-linear relationships between ingredients[7].

3. Support Vector Machines (SVM): SVM is a linear or non-linear algorithm that finds a hyperplane that maximally separates classes. In the context of a recipe suggestion website, SVM could be used to classify recipes into categories such as vegetarian or non-vegetarian based on their ingredients. However, SVM may not perform as well as Random Forest for this task since it may struggle with classifying recipes that have similar ingredient lists but belong to different categories[4].

4. Decision Trees: Decision Trees are a simple yet powerful algorithm that partitions data into smaller subspaces based on the values of their features. In the context of a recipe suggestion website, Decision Trees could be used to suggest recipes based on the presence or absence of specific ingredients. However, Decision Trees may not perform as well as Random Forest for this task since they tend to overfit to the training data and may not generalize well to new data.

5. GPT-2: GPT-2 is a transformer-based language model that can be used for recipe generation and recommendation by training it on a large corpus of recipe data. The model has the ability to capture complex relationships between ingredients and generate natural language text, making it suitable for tasks like recipe generation and recommendation. Additionally, GPT-2 has achieved state-of-the-art results on a range of natural language processing tasks, including language modeling, text completion, and question answering, making it a promising choice for recipe suggestion websites[1].

7. Conclusion and Future Scope

To conclude, transformers such as GPT-2 offer significant advantages over traditional machine learning algorithms like KNN, SVM, decision trees, and random forest for recipe generation models. With their ability to handle large amounts of unstructured text data and generate human-like responses, transformers can provide more personalized and engaging recipe recommendations to users based on their preferences and previous interactions with the website. Additionally, transformers can learn complex relationships between words and phrases and produce more creative and diverse outputs than rule-based or statistical models. Transformers can also generate new and innovative recipe ideas based on the ingredients provided, making the website stand out from its competitors. Furthermore, transformers can be fine-tuned on specific recipe datasets to improve their performance and accuracy, making them highly adaptable to different types of recipe data. When selecting an algorithm, it's essential to consider factors such as the size of the dataset, the number of features, the type of data, and the desired outcome of the task. Therefore, it's crucial to thoroughly evaluate the pros and cons of different algorithms and select the one that is best suited for the specific task at hand.

A basic recipe generation model developed using transformers and GPT-2 has vast potential for future development and improvement. One potential area for expansion is multimodal recipe generation, which could incorporate additional modalities such as images or user preferences to produce more contextually relevant and personalized recipe recommendations. Real-time feedback and adaptation could also be incorporated to optimize recipe recommendations based on user interactions. Additionally, integration with e-commerce platforms would provide a seamless end-to-end solution for users by allowing them to purchase necessary ingredients directly from the website. Lastly, customized recipe generation can be achieved by training the model on user-specific data such as dietary restrictions, cuisine preferences, or cooking skill level, which can result in even more personalized and relevant recipe recommendations.

Acknowledgement

We would like to express our gratitude to our project guide, Prof. Mrs. Dipalee Rane, for providing us with guidance, supervision, and necessary information to complete this research paper. We also appreciate the support from the Head of Department of ComputerEngeneering, Dr. Mrs.M.A.Potey, for her cooperation and encouragement throughout the completion of this research paper.

REFERENCES

1. MOGAN GIM , DONGHYEON PARK , MICHAEL SPRANGER , (Member,IEEE), KANA MARUYAMA2 , AND JAEWOO KANG, 14 Oct 2021 IEEE RecipeBowl : A Cooking recommender for ingredients and recipes using set-transformer.

2. Nilesh, Dr. Madhu Kumari, Pritom Hazarika, Vishal Raman, " 2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW) " Recommendation of Indian Cuisine Recipes based on Ingredients

3. Suyash Maheshwari, Manas Chourey, "International Research Journal of Engineering and Technology (IRJET)" Recipe Recommendation System using Machine Learning Models.

4. Srikar Amara, R. Raja Subramanian ; 2020 6th International Conference on Advanced Computing and Communi- cation Systems (ICACCS), Collaborating personalized recommender system and content-based recommendder system using TextCorpus.

5. Shobhna Jayaraman'. Tanupriya Choudhury, Praveen Kumar; 2017 International Conference On Smart Technology for Smart Nation; Analysis of Classification Models Based on Cuisine Prediction Using Machine Learning.

6. 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICACCCT) 1857 ISBN: 978-1-6654-3811-7/21/31.00 ©2021 IEEE Recipe Recommendation System with Ingredients Available on User; Siddharth Raj , Dr Ajay Shanker Singh, Ayush Sinha, Anandhan K., Mayank Srivastav.

7. 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), A Novel recipes recommendation system Based on Knowledge-Graph ; BoHuang*, Xiaonan Shi, Rongqiang Wang, Chenyang Wang, Yuanhao Han.

8. S. Praveen, M.V. Prithivi Raj, R. Poovarasan, V. Thiruvenkadam, M. Kavin Kumar 5; International Research Journal of Engineering and Technology (IRJET) Volume: 06 Issue: 02 — Feb 2019 Discovery of Recipes Based on Ingredients using Machine Learning.

9. Naoki SHINO, Ryosuke YAMANISHI, Junichi FUKUMOTO 2016 5th IIAI International Congress on Advanced Applied Informatics, Recommendation System for Alternative-ingredients Based on Co-occurrence Relation on Recipe Database and the Ingredient category.

10. 2019 IEEE 2nd International Conference on Information Systems and Computer Aided Education (ICISCAE) , Collaborative Filtering Recommendation Algorithm Based on Random Forest Filling, Zhongwei Wang , HangpingQiu, Yi Sun, Qiaoyu De.