Drug Review Analysis By Multifaceted ML Techniques

By
Chethana Kadirimangalam

Table of Contents

Introduction

Drug Review Analysis is a crucial area of study in the field of healthcare and pharmaceuticals. With the widespread availability of online platforms where users share their experiences with various medications, there exists a wealth of valuable data that can be analyzed to gain insights into the effectiveness, side effects, and overall sentiment surrounding different drugs. In this project, we delve into the realm of drug review analysis, aiming to extract meaningful patterns and information from a large dataset of drug reviews.

Problem Statement

Understand Effectiveness: Determined the effectiveness of different drugs in treating specific medical conditions based on user reviews.

Identify Side Effects: Identified common side effects associated with different medications to aid in better understanding their safety profiles.

Explore Sentiment: Analyzed the sentiment of drug reviews to gauge overall user satisfaction and perception of different medications.

Model Building: Implemented machine learning and deep learning models to classify drugs effectiveness based on reviews and ratings.

Dataset loading and Overview

- The dataset is extracted from UCI Machine Learning Repository by using pandas library.
- The size of the training dataset is printed, indicating it contains 161,297 rows and 7 columns.
- Column names are displayed to understand the information contained in each column.
- The first 5 rows of the training dataset are shown to provide a glimpse of the data.
- Summary statistics of numeric columns are computed, including count, mean, standard deviation.

Data Pre-Processing

Cleaning Missing Values

Missing values in the 'condition' column are replaced with the mode of the column.

Handling Text Data

Removing stopwords to eliminate common, less informative words.

Removing punctuation marks to streamline the text.

Tokenization to break down the text into individual words for analysis.

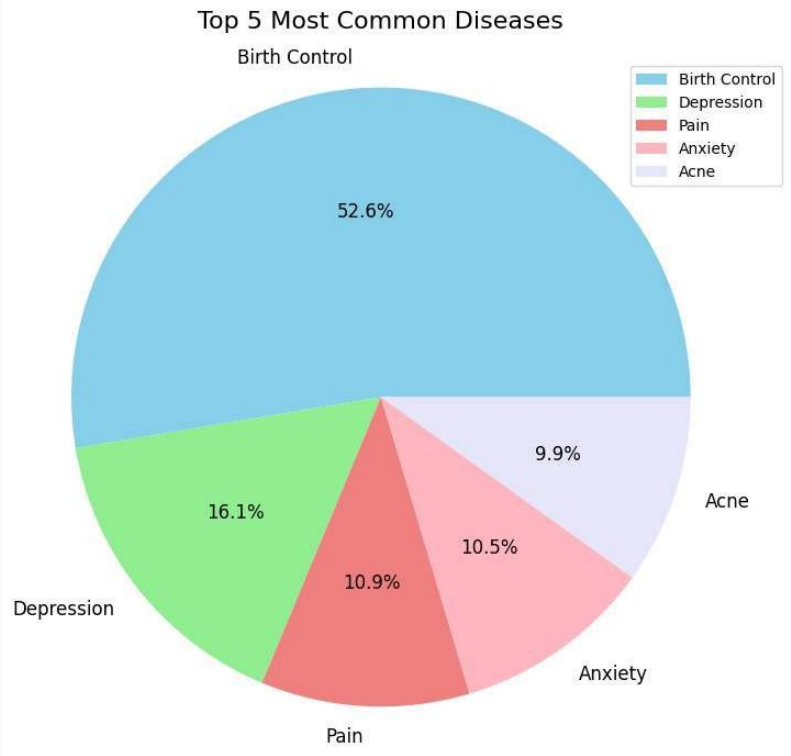
Encoding Categorical Variables

Categorical variables column in the dataset, are encoded into numerical format for machine learning models to understand.

TF-IDF Vectorization

TF-IDF vectorization is performed on the cleaned text data using the TfidfVectorizer from the sklearn.feature_extraction.text module.

Exploratory Data Analysis



52.6% **Birth Control**

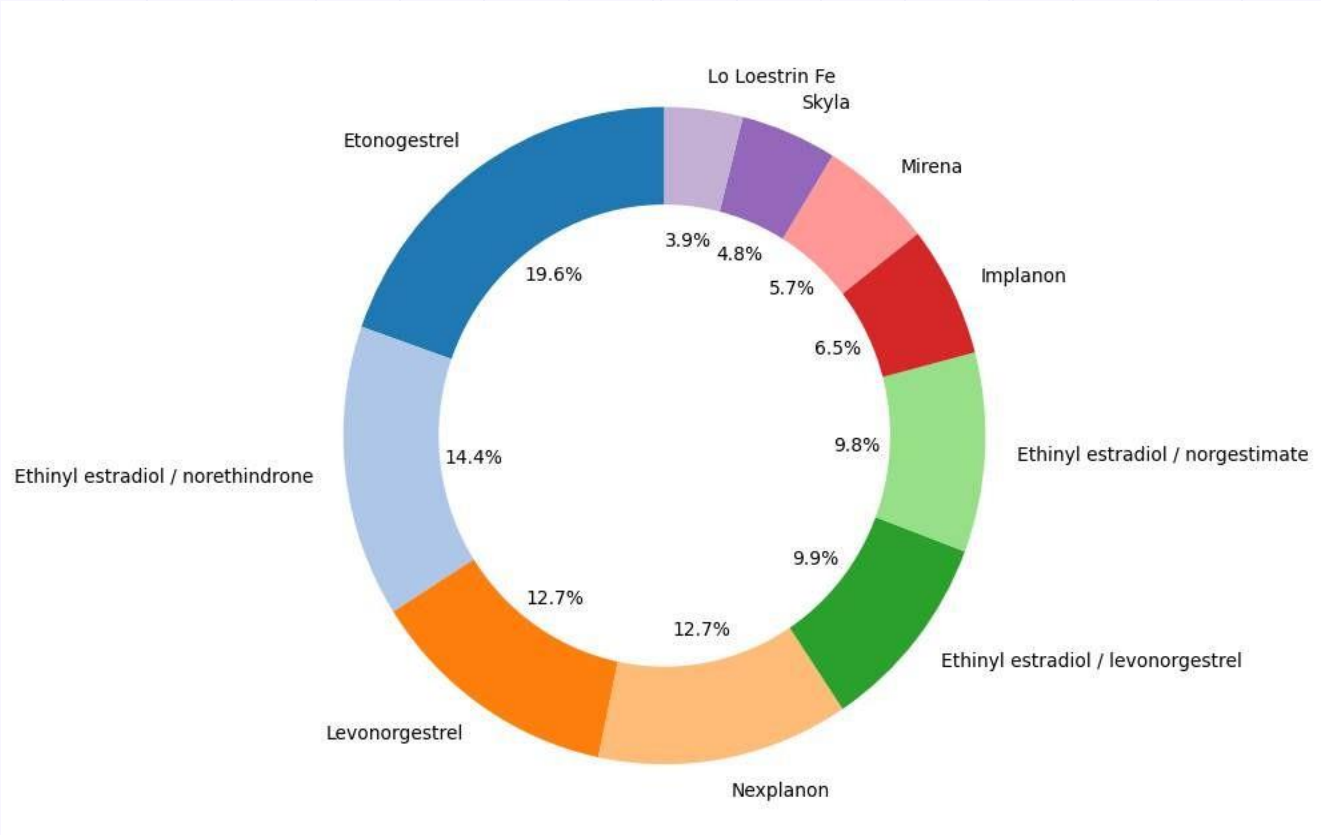
16.1% **Depression**

10.9% **Pain**

10.5% **Anxiety**

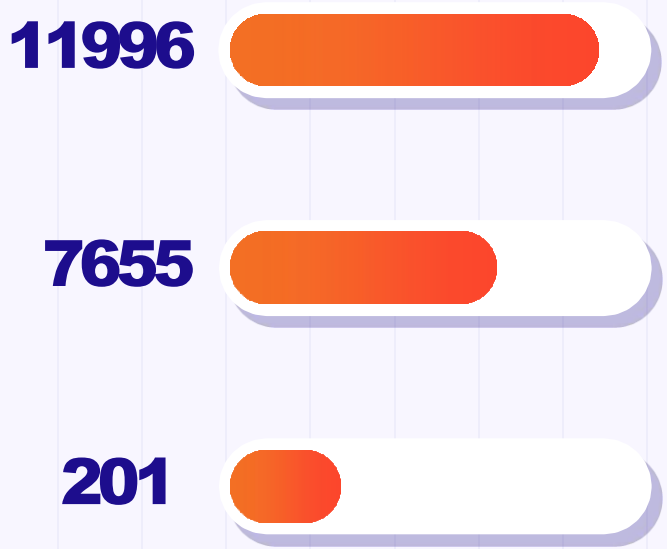
9.9% **Acne**

Total number of unique conditions: 884



Total number of unique drugs for “Birth Control”: 417

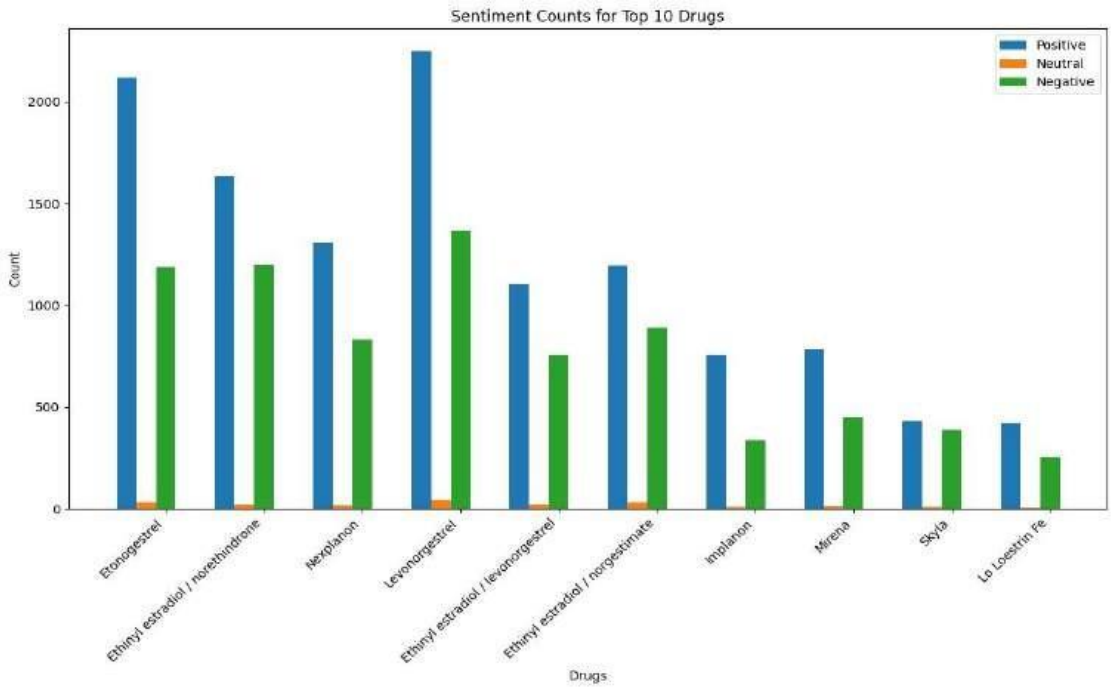
Sentiment Analysis



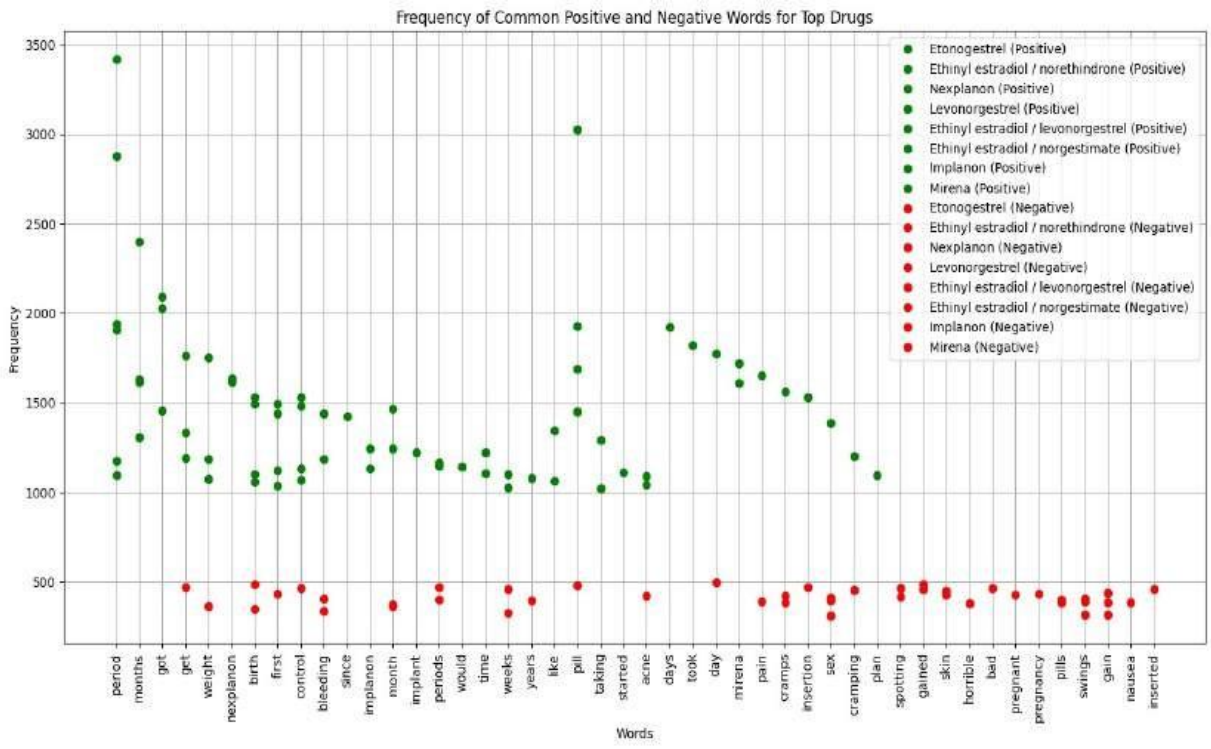
- We conducted sentiment analysis on reviews related to birth control, which accounted for 29,687 reviews in total.
- Utilized TextBlob for sentiment analysis, categorizing reviews as positive, negative, or neutral based on polarity.

```
Sentiment Counts for Each Drug:
Drug: Etonogestrel
Positive: 2119, Negative: 1186, Neutral: 31

Drug: Ethinyl estradiol / norethindrone
Positive: 1632, Negative: 1198, Neutral: 20

Drug: Nexplanon
Positive: 1307, Negative: 832, Neutral: 17

Drug: Levonorgestrel
Positive: 2246, Negative: 1367, Neutral: 44

Drug: Ethinyl estradiol / levonorgestrel
Positive: 1106, Negative: 758, Neutral: 24

Drug: Ethinyl estradiol / norgestimate
Positive: 1196, Negative: 891, Neutral: 30

Drug: Implanon
Positive: 755, Negative: 336, Neutral: 11

Drug: Mirena
Positive: 781, Negative: 447, Neutral: 14

Drug: Skyla
Positive: 431, Negative: 386, Neutral: 7

Drug: Lo Loestrin Fe
Positive: 423, Negative: 254, Neutral: 3
```

Performs sentiment analysis for each of the top 10 drugs separately.
Counts the number of reviews with positive, negative, and neutral sentiments for each drug.



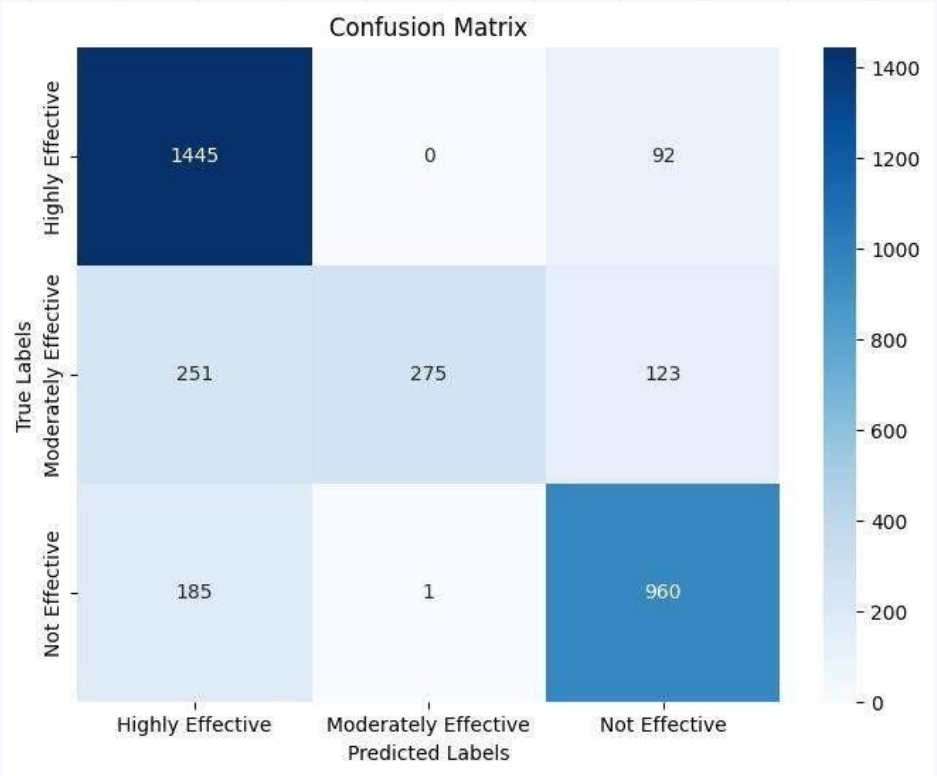
- Word clouds are generated to illustrate the most common words in both positive and negative reviews for each drug.
- Word frequencies are analyzed to determine the most frequent words for top 10 drug. This analysis offers insights into the language used in reviews and aids in understanding the sentiments associated with each drug, facilitating informed decision-making and further exploration of user experiences.



Random Forest Classifier

- **Model Selection**: Random Forest Classifier is chosen as the classification algorithm due to its ability to handle high-dimensional data and provide accurate predictions.
- **Model Training**: The classifier is trained on the TF-IDF transformed training data. During training, multiple decision trees are built on different subsets of the data and features, adding randomness and reducing overfitting.
- **Model Evaluation**: Performance metrics such as accuracy, precision, recall, and F1-score are calculated to assess the model's effectiveness.

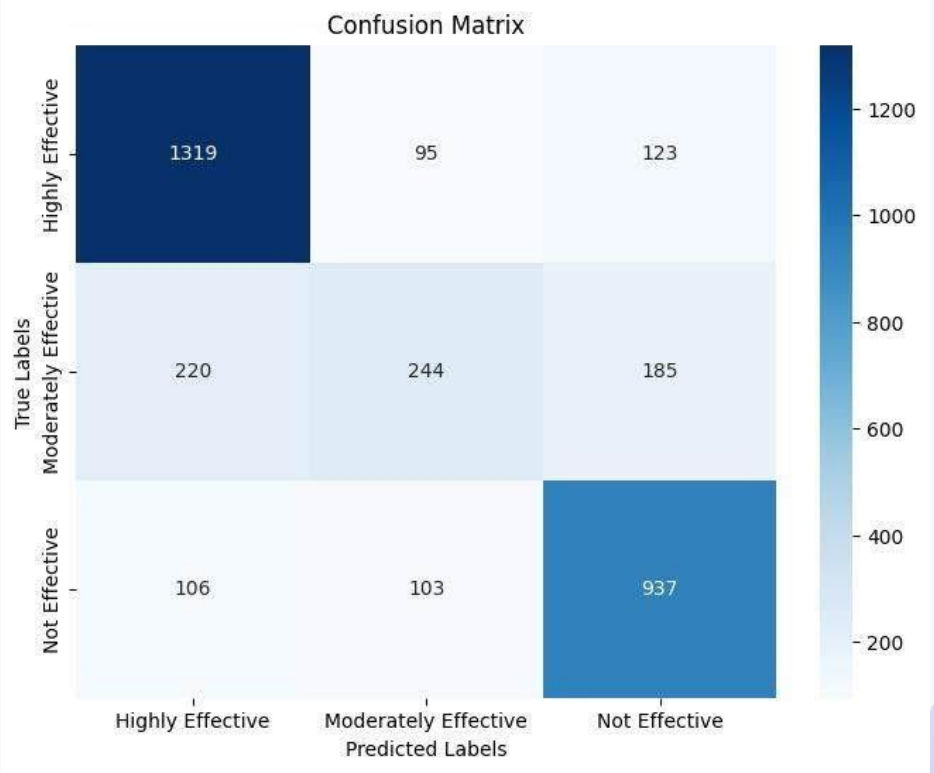
  Accuracy: 80.4%
  Precision: 82.9%
  Recall: 80.4%
  F1 Score: 79%



Stacking Classifier

- **Model Initialization**: Initialized base models (Logistic Regression, Naive Bayes) with hyperparameter tuning using GridSearchCV to optimize their performance..
- **Stacking Model Construction**: A meta-classifier (Random Forest) is used to combine predictions from base models. StackingClassifier aggregates predictions from base models and learns to predict the final outcome.
- **Model Training**: Train base models and stacking model on TF-IDF transformed training data.
- **Model Evaluation**: The stacking model is evaluated using classification metrics such as precision, recall, and F1-score.

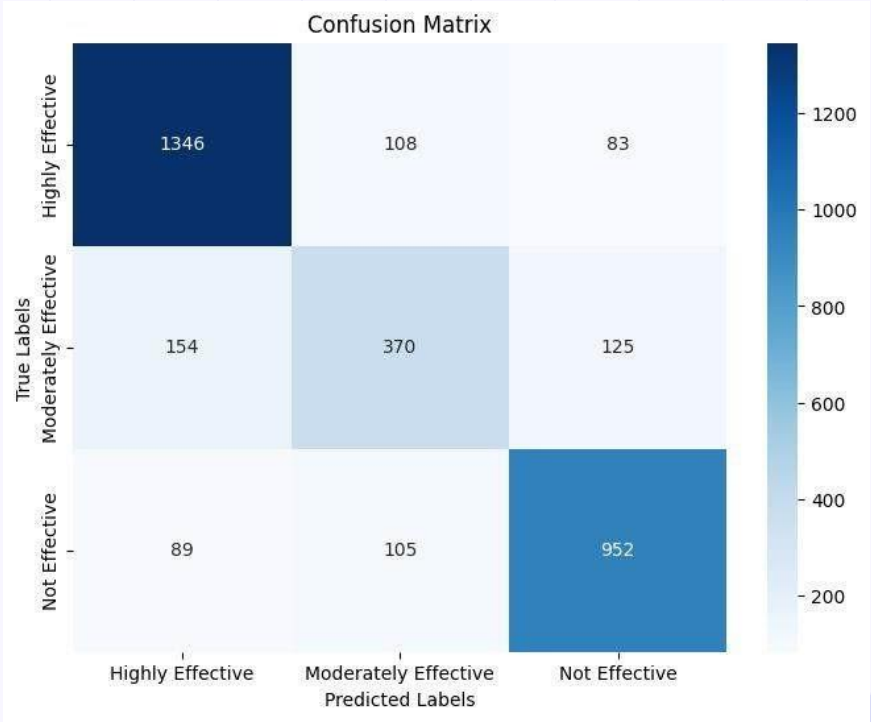
Accuracy: 74.9%
Precision: 73.4%
Recall: 74.9%
F1 Score: 73.8%

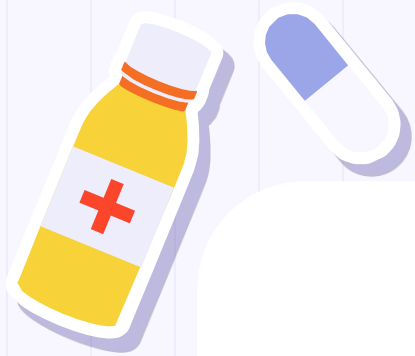


Dense Neural Network (DNN) model

- **Model Definition**: A Dense Neural Network (DNN) model is constructed using Keras Sequential API.
- **Model Construction**: Sequential Deep Neural Network (DNN) with three layers:
  - Input layer (128 neurons)
  - Hidden layer (64 neurons)
  - Output layer (3 neurons for rating categories)
- **Compilation**: The model is compiled with 'adam' optimizer and 'sparse_categorical_crossentropy' loss function.
- **Model Training and Evaluation**: Training progress is monitored with accuracy, loss, and validation accuracy metrics.

Accuracy: 80%
Precision: 79.6%
Recall: 80%
F1 Score: 79.8%





Conclusion

Leveraging advanced analytics in drug review analysis through machine learning and deep learning techniques, we gained valuable insights into drug effectiveness, side effects, and user sentiment from a vast dataset of drug reviews. This analysis enhances our understanding of medication usage and aids in decision-making for healthcare professionals and pharmaceutical stakeholders.


A picture is worth a thousand words