**Summary**

**Introduction**

Train and test datasets include benefits\_review, side\_effects\_review, comments\_review, rating columns.

**Methodology**

The text model is constructed by using Replacement blanks and stop words, Tokenization, Part-of-Speech Tagging, Named Entity, Lemmatization and TF-IDF Vectorizer. At the same time, the text model has the best performance by selecting the number of most important features in TF-IDF Vectorizer and machine learning algorithms, which include SVM, Logistic Regression, Random Forest, Extra Trees, KNN and Naïve Bayes.

**Result**



Table 1 Using top 1000 features in the text model

When building models using KNN and Naïve Bayes machine learning algorithms, their accuracy and F1 score are far lower than other algorithms, the accuracy and F1 score of Extra Trees are the same as Random Forest. So, using SVM, Logistic Regression and Random Forest to predict in train dataset and compare their results. From the form above, we can see that model performance can be improved by using all columns or side\_effects\_review or side\_effects\_review + comments\_review or benefits\_review. Then, use all the features to see if it can improve the accuracy of the text model.



Table 2 Using all features in the text model

After computing their accuracy and F1 score, using all features and SVM machine learning algorithm in the text model can get the best performance.

Finally, using all features and SVM machine learning algorithm to predict test dataset, the accuracy is 99.90%, and F1 score is 99.91%.