# Predict the relevance of search results using traditional and advance techniques

Hussnain Khalid (huskh803)

December 25, 2022

**Abstract**

This research is about the performance of search engine algorithms that use different NLP methods in order to give more relevant results for a specific query. Different traditional machine learning models and some neural network models are used to measure the relevance of search results. In this research, such models were applied to query based on product title and product description to predict median relevance. Performance analysis is based on accuracy and model complexity.

## 1 Introduction

Today the ecommerce business is the fastest growing business, and the amount of data is growing exponentially, it becomes more and more important to be able to find relevant information. Search engines are designed to fetch information relevant to search queries. A query is used based on algorithms to fetch optimized search results.

Search engine performance is most important for ecommerce business. In order to be efficient different techniques are used to fetch desired data. But it's important to focus on the quality of data instead of quantity of the data.

Natural language processing (NLP) is used to research on how to make computers understand human language. As the use of search engines and the amount of processing data increases it becomes very important to handle search queries. Today search engines uses NLP for lemmatization of words in a query to fetch related result to original keywords. So, it's important to measure the relevance of search results to make the search engine more powerful and efficient.

One of the important task for this research is also to find out how advanced methods like neural networks can help to predict the relevance of search results.

# 2 Theory

## 2.1 Multinomial Naive Bayes

Naive Bayes Classifier is a probabilistic algorithm based on Bayes' theorem with the "naive" assumption of conditional independence between every pair of a feature. This model was used frequently during the course but the reason for mentioning and using this model is that it performed much better than other models with just some data cleaning as data cleaning was the biggest hurdle in this implementation.

## 2.2 Support Vector Machine (SVM)

SVM is another basic model that determines the best decision boundary between vectors that belong to a given group and vectors that do not belong to it. SVM is used because it proves to be a very effective model in this case[1]. It works almost the same as Naive Bayes Classifier but there are still few differences which will be explained later.

## 2.3 Random forest

Random forest module from sklearn is implemented, as a bagging method, it generates a number of trees instead of a single tree. The training procedure is shown below.

1. The number of trees(estimators) in the forest (10,50,100).
2. The maximum depth of the tree (4).
3. The minimum number of samples required to split (10).

Grid search is used to find the best estimator and used but still the results are not better than other traditional methods. It became useless as our data is quite biased (reason will be discussed in the Method section).

## 2.4 Multi-layer Perceptron regressor (MLP)

MLP[2] is a neural network that has 3 or more layers of perceptrons (input layer, hidden layer and output layer) [3]. It's a single direction model. It uses back-propagation technique to receive feedback on error and adjust the weights accordingly.

[1] https://medium.com/@bedigunjit/simple-guide-to-text-classification-nlp-using-svm-and-naive-bayes-wit

[2] https://www.geeksforgeeks.org/difference-between-multilayer-perceptron-and-linear-regression/

[3] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html

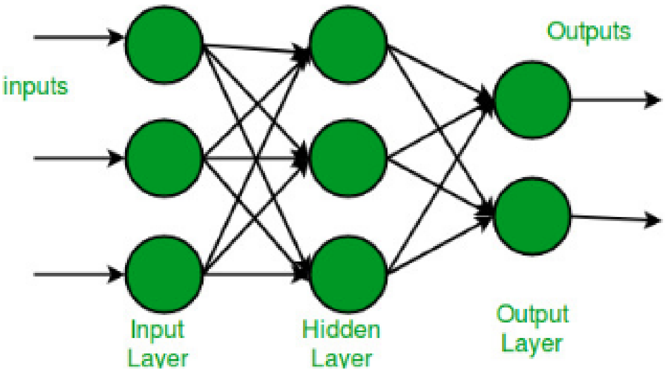


Figure 1: MLP topology

### 2.5 Extra-Trees Regressor (ET)

ET[4] is an ensemble machine learning approach very similar to Random Forest but there are few differences, like Random Forest uses bagging technique to select variation of training data on the other hand ET uses entire dataset for training. Extra Trees are much faster as compared to Random Forest with respect to computational cost[5].

### 2.6 Elastic-Net

Elastic-Net is a linear regression model which uses the penalty factor from Lasso and Ridge regression for regularization[6]. The benefit of using Elastic-Net is it keeps both penalties balanced, for better performance.

## 3 Data

### 3.1 Raw Data

Data is taken from a Kaggle competition[7]. The size of the data is 10K rows and 5 columns showed below

1. Id
2. Query - Query text.
3. Product_title - Product title text.
4. Prodcut_description - Product description text.

[4] https://orbi.uliege.be/bitstream/2268/9357/1/geurts-mlj-advance.pdf
[5] https://towardsdatascience.com/what-when-how-extratrees-classifier-c939f905851c
[6] https://machinelearningmastery.com/elastic-net-regression-in-python/
[7] https://www.kaggle.com/competitions/crowdflower-search-relevance/overview

5. Median_relevance - Median relevance score.

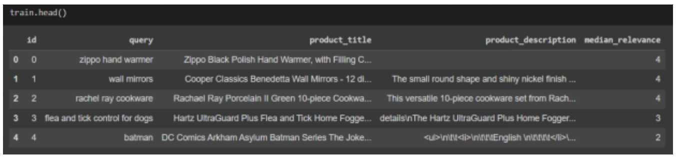
The Median_relevance column contains 4 different classes [4,3,2,1], where 4 is for most relevant product according to the query and 1 is for least relevant product.

### 3.2 Data Cleaning

Data contains 2444 null values in the product_description column. As missing values are only in the product_description column we will replace them with empty string (the purpose is explained in the Method section). Data is split into training and test data randomly, where training data is 75% and test data is remaining 25%. Id column is dropped in most cases as its useless for us mostly.

### 3.3 Data Example

Training data:

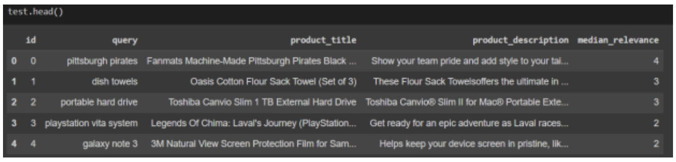


```
train.head()
```

| | id | query | product_title | product_description | median_relevance |
|---|---|---|---|---|---|
| 0 | 0 | zippo hand warmer | Zippo Black Polish Hand Warmer, with Filling C... | | 4 |
| 1 | 1 | wall mirrors | Cooper Classics Benedetta Wall Mirrors - 12 di... | The small round shape and shiny nickel finish ... | 4 |
| 2 | 2 | rachel ray cookware | Rachael Ray Porcelain II Green 10-piece Cookwa... | This versatile 10-piece cookware set from Rach... | 4 |
| 3 | 3 | flea and tick control for dogs | Hartz UltraGuard Plus Flea and Tick Home Fogge... | details\nThe Hartz UltraGuard Plus Home Fogger... | 3 |
| 4 | 4 | batman | DC Comics Arkham Asylum Batman Series The Joke... | <ul>\n\t\t<li>\n\t\t\tEnglish \n\t\t\t\t</li>\... | 2 |

Figure 2: Training Data

Test data:



```
test.head()
```

| | id | query | product_title | product_description | median_relevance |
|---|---|---|---|---|---|
| 0 | 0 | pittsburgh pirates | Fanmats Machine-Made Pittsburgh Pirates Black ... | Show your team pride and add style to your tai... | 4 |
| 1 | 1 | dish towels | Oasis Cotton Flour Sack Towel (Set of 3) | These Flour Sack Towelsoffers the ultimate in ... | 3 |
| 2 | 2 | portable hard drive | Toshiba Canvio Slim 1 TB External Hard Drive | Toshiba Canvio® Slim II for Mac® Portable Exte... | 3 |
| 3 | 3 | playstation vita system | Legends Of Chima: Laval's Journey (PlayStation... | Get ready for an epic adventure as Laval races... | 2 |
| 4 | 4 | galaxy note 3 | 3M Natural View Screen Protection Film for Sam... | Helps keep your device screen in pristine, lik... | 2 |

Figure 3: Test Data

# 4 Method

So, in order to evaluate the results, I have to choose a baseline and, in this case, I have to beat 60% accuracy. I chose 60 because in the data the relevance of search query with perfect match is around 60% (median_-relevance == 4).

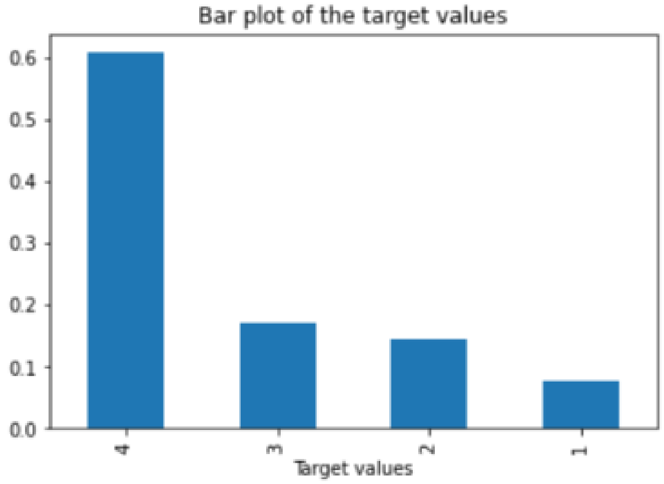


Figure 4: Target Value

We can also call it a dummy classifier which is usually built using the most frequent scheme.

Before using any method first, I implement a few techniques to set up training and test data for the model. To do so I dropped the **Id** column for traditional machine learning models and also combined **product_title** and **product_description** columns together as 1 column so we don't have to drop null rows because of null values present in the product_description column. Also, this can help to evaluate the vocabulary better.

For preprocessing different techniques are used to remove stop words and some specific stop words added in the list for removal as they are useless for the model training. Also the concept of stemming is used to deal with morphological variants of words. TfidfVectorizer is used to transform the test to feature vectors that is used as an input to our estimators. Pipeline from Sklearn is also used to connect all steps together for different parameters.

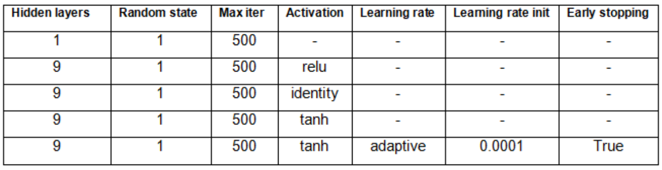
After all that new training and test data frames are set for implementation of models. And the first traditional and the basic model for this task I used is **Naive Bayes** which is the most common model and gives an expected good result too.

**SVM** proves to be also a nice classifier in our case there are few differences which will be explained in the result section. I used different penalty parameters of error term but C=10 proves to be best of all these. There is no way to know the best value for C. It only depends upon what result you will have after applying on test data.

**Random Forest Classifier** is not up to the mark and performed really badly. Though different parameters were applied like max_depth = 4, min_samples_leaf = 10 and different estimators for number of trees (10, 50, 100) are used with a grid search approach for the best estimator.

Other than these traditional methods some advanced complex methods are also applied to check how they react but all of them performed really badly. In the discussion section I'll explain what could be the reason for that according to my experiments.

**Multi-layer Perceptron** classifier is used with different setups shown below.



| Hidden layers | Random state | Max iter | Activation | Learning rate | Learning rate init | Early stopping |
|---|---|---|---|---|---|---|
| 1 | 1 | 500 | - | - | - | - |
| 9 | 1 | 500 | relu | - | - | - |
| 9 | 1 | 500 | identity | - | - | - |
| 9 | 1 | 500 | tanh | - | - | - |
| 9 | 1 | 500 | tanh | adaptive | 0.0001 | True |

Figure 5: MLP Training Summary

In the last configurations activation function 'tanh' proves to be best for this task, 'adaptive' learning rate was used to keep the learning rate constant as long as training loss keeps decreasing and early stopping is also used to terminate training when validation score is not improving [8].

**Extra-Trees regressor** with following configuration is used n_estimators = 250 and min_samples_split = 10 means number of trees in the forest are 250 and minimum number of samples required to split is 10.

**Elastic Net**[9] is used with these configurations alpha = 0.01, l1_ratio = 0.5 and fit_intercept = False. Where alpha is the constant multiplies the penalty term. With l1_ration = 0.5 we can use both L1 and L2 penalty combination and fit_intercept is set to be false as it is used to check if the data is already centered or not.

All these advanced methods MLP, Extra-Trees and Elastic Net are used as ensemble learning[10], where their predictions are combined together to improve the results.

## 5 Result

As we already talked about baseline setup according to median_relevance which we took 60%. Now we can compare the results for all these traditional and advanced techniques for prediction of search result relevance. For **traditional methods** we will see different techniques to check how the models react for median_relevance based on precision, recall and f1 score as accuracy of all these models are around 62%. As precision provides us

[8] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

[9] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html

[10] https://scikit-learn.org/stable/modules/ensemble.html

the information that how precise is the model to predict right/positive, how many of them are actually right/positive.

$$\text{Precision} = \frac{\textit{True Positive}}{\textit{True Positive}+\textit{False Positive}}$$

$$= \frac{\textit{True Positive}}{\textit{Total Predicted Positive}}$$

Recall on the hand is a very important measure if the cost of false negatives is very high. But in our case, we can rely on precision as it is the most important factor for us to compare relevance prediction study.

$$\text{Recall} = \frac{\textit{True Positive}}{\textit{True Positive}+\textit{False Negative}}$$

$$= \frac{\textit{True Positive}}{\textit{Total Actual Positive}}$$

And for F1 score it is a function of precision and recall. We normally use it when we want a balance between precision and recall. Though precision is our most important factor.

$$\text{F1} = 2 \text{ x } \frac{\textit{Precision}*\textit{Recall}}{\textit{Precision}+\textit{Recall}}$$

## 5.1 Naive Bayes

| Classification | Precision | Recall | F1 score |
|---|---|---|---|
| 1 | 0.57 | 0.24 | 0.34 |
| 2 | 0.33 | 0.24 | 0.28 |
| 3 | 0.31 | 0.20 | 0.24 |
| 4 | 0.70 | 0.86 | 0.77 |

### 5.2 SVM

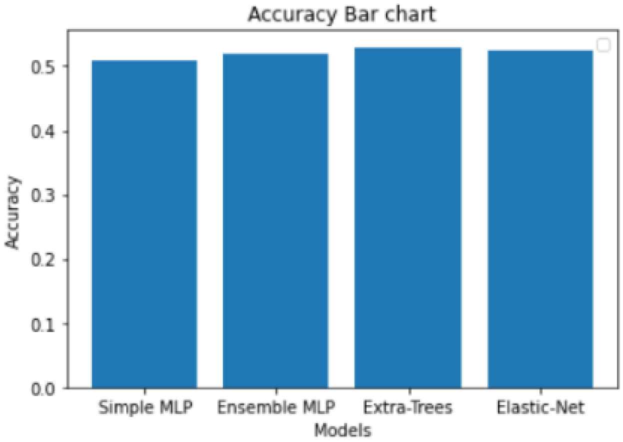
| Classification | Precision | Recall | F1 score |
|---|---|---|---|
| 1 | 0.50 | 0.05 | 0.09 |
| 2 | 0.38 | 0.11 | 0.17 |
| 3 | 0.41 | 0.09 | 0.14 |
| 4 | 0.64 | 0.95 | 0.76 |

### 5.3 Random Forest

| Classification | Precision | Recall | F1 score |
|---|---|---|---|
| 1 | 0.00 | 0.00 | 0.00 |
| 2 | 0.00 | 0.00 | 0.00 |
| 3 | 0.00 | 0.00 | 0.00 |
| 4 | 0.62 | 1.00 | 0.76 |

For **advanced methods** we will use accuracy measures to check how well they react as their accuracies are very low so we can't use any other measures. **MLP** with the best configurations gives the accuracy of 50.8%. Many other configurations are also applied but still the result is not more than 50%.

That's why different methods are also applied but their results are almost the same so at the end all are combined together using ensemble method. The conclusion will be discussed in later section. MLP here gives accuracy of around 52%, **Extra-Trees** gives accuracy of 53% and Elastic-Net combined with them gives **52.3%** accuracy only.



## 6 Discussion

First of all, we will focus on traditional methods applied to predict relevance of search results, we set the baseline to 60% described before. We got accuracy of 62% from all the traditional methods. That's why we chose a different technique to check which method performed well.

We choose to go with precision, recall and F1 score and for us precision is the most important measure described before in the results section. We can see Naïve Bayes performing really well as the precision of median_relevance = 4 is 70% compared to other methods where we get 64% with SVM and 62% with Random Forest. As mentioned before Random Forest performed worst in all these methods because the precision score for other classes is 0% which makes sense as the data is very biased for class 4.

It can be seen that unbiased data is a limitation and may be with biased data results will be very different. But having high numbers for class 4 is much better which proves that the search algorithm is quite good.

Another possibility which proves very handy is to clean the data as much as possible, the cleaner data we have the better the models work. After using normal text mining techniques like removing stop words, special symbols and numeric data we were still missing a few specific words which are useless for the model so they were added separately in the stop word dictionary to have better results. It's really hard to find these kinds of words from the data. It proves that more understanding and exploration of data is needed.

For the bad performance of advanced methods data cleaning can be an issue and also size of data which is always a barrier for deep neural methods. More study of data can be proved handy to assign proper weights to classes. Which needs more time and exploration of data.

## 7 Related work

In this thesis project[1] Mobyen Uddin Ahmed investigates the performance of search engine that uses NLP methods in order to find the use of triplets could improve search results. In this paper[2] Rodrigo Nogueira, Wei Yang, Jimmy Lin and Kyunghyun Cho proposed a method to improve the information retrieval effectiveness of a search engine by expanding documents based on neural networks. Also, a lot of work is done in this Kaggle competition [11] the goal was to develop an open-source model that measure the relevance of search results to help small business owners to compete against resource rich competitors.

## 8 Conclusion

Though deep learning is very popular and is a breakthrough. But still traditional methods are better in a lot of cases, especially when we have limited data. This whole implementation and research prove that it's not

[11] https://www.kaggle.com/competitions/crowdflower-search-relevance/overview

a better approach to dig too deep when the cost of action is higher and outcome is not much. Also, measuring performance of models shouldn't only rely on accuracy of the model specially when dealing with highly biased data. At the end Exploratory Data Analysis (EDA) plays the most important role to train a better model.

## References

[1] Carlstedt Martin. Using nlp and context for improved search result in specialized search engines. 2017.

[2] Rodrigo Frassetto Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. Document expansion by query prediction. *CoRR*, abs/1904.08375, 2019.