IMDB Review Rating - Sentiment Analysis

Balassubramanian Srinivasan and Vikas Velagapudi

Video Link : CS 657 Final Project Video Presentation (https://youtu.be/Jh5hZaE31JA)

Introduction

This project explores the valuable insights hidden within the vast array of reviews on IMDb, a premier database for films, TV shows, and similar content. The project's cornerstone is a comprehensive dataset obtained from IMDb, via Kaggle, containing detailed information on individual reviews. These reviews, contributed annually by a vast community, are more than just opinions; they contain crucial data points that influence viewer choices and perceptions.

In this project, we analyzed a dataset, comprising elements like review ID, reviewer username, movie title, rating, review summary, date, spoiler tag, review detail, and a helpfulness score. Initially in JSON format, the data was converted to CSV for ease of analysis, with a key preprocessing step being the creation of a 'Sentiment' column based on ratings, categorizing them as negative (1-5) into 0 or positive (6-10) into 1. Our analysis involved developing and cross-validating multiple machine learning models—Logistic Regression, Linear SVC, Random Forest, and MLP—to determine the most effective. Additionally, we innovatively used the 'Helpfulness' metric to identify the most valuable reviews and reviewers in the IMDb community.

About the Dataset

The dataset is an aggregation of publicly available individual data from IMDb. It encompasses various elements, each represented in a structured format:

- Review ID: A unique identifier assigned by IMDb.
- Reviewer: The public identity or username of the reviewer.
- Movie: The title of the film or TV series being reviewed.
- Rating: A numerical score out of 10, with older reviews possibly lacking this data.
- Review Summary: A concise encapsulation of the reviewer's opinion.
- Review Date: The timestamp of when the review was posted.
- Spoiler Tag: A binary indicator where 1 represents a spoiler, and 0 signifies its absence.
- Review Detail: An in-depth account of the reviewer's perspective.
- Helpfulness: A measure of the review's utility, expressed as Upvotes and Downvotes.

Architecture

Data preprocessing

The architecture of our data processing pipeline begins with the conversion of the IMDb dataset from JSON to CSV format, utilizing Apache Spark for its robust handling of large datasets. The process is initiated by reading the JSON file with a predefined custom schema to structure the data effectively. In the interest of simplification and relevance, we remove the 'review_date' column from our DataFrame.

To facilitate sentiment analysis, a new column 'sentiment' is created, using a threshold value of 5 to distinguish between negative (ratings below 5) and positive (ratings above 5) sentiments. This classification is crucial for subsequent analytical models. We also transform the 'helpful' array into a string format, concatenating its elements for a more streamlined representation. After these transformations, the final DataFrame is written to a CSV file, with headers included for clarity, and stored in a designated location. This streamlined and efficient conversion process forms the foundation of our analytical pipeline, ensuring data integrity and ease of access for machine learning models.

Converting Words to Tokens and Vectors

In this phase, we focused on preparing the 'review_summary' column for machine learning training by encoding, tokenizing, and converting text into vectors. This process involves several steps, implemented using Apache Spark's ML library. Firstly, we filter out records with null values in 'review_summary' and 'rating' to maintain data integrity. We then tokenize the review text into individual words. The next step involves removing common stopwords to enhance the quality of text data for analysis. Subsequently, we employ the Word2Vec algorithm to convert these filtered words into numerical vectors, which is a critical step for enabling machine learning algorithms to process textual data. Word2Vec is a technique in natural language processing that converts words into numerical vectors, capturing semantic relationships and contexts in a multi-dimensional space. In parallel, we converted the 'rating' column into a binary 'sentiment' value (1 for positive, 0 for negative) based on a threshold. This entire sequence is structured as a pipeline in Spark, ensuring an efficient and systematic transformation of text data into a format suitable for advanced analytical models. This transformed data is then used in training various machine learning algorithms implemented.

Machine Learning Models

Cross-Validation for models

Cross-validation is a crucial technique in machine learning, used for evaluating the effectiveness of a model's performance. In the provided code snippet, a CrossValidator is implemented, indicating an advanced approach to model training and evaluation. This process involves setting up a pipeline and parameter grid, followed by specifying an evaluator and the number of folds, in this case, 10. This means the data is divided into ten parts, with each part in turn used as a test set and the rest as training data. Additionally, a seed is set for reproducibility, ensuring consistent results across different runs. This is used across the four machine learning models that we built.

1. **Logistic Regression**

   Logistic Regression (LR) for sentiment analysis involves a multi-step process, tailored to optimize model performance using Apache Spark's ML library. Initially, we define the LR model, specifying 'word_vectors' as the feature column and 'sentiment' as the label column. This model is then integrated into a pipeline, streamlining the process flow. For evaluation, we employ the BinaryClassificationEvaluator with the Area Under the Receiver Operating Characteristic (ROC) curve (AUC-ROC) as the metric, which effectively measures the model's

ability to distinguish between classes. The core of our approach lies in hyperparameter tuning, achieved through a ParamGridBuilder. We experiment with different values of the regularization parameter ('regParam') and ElasticNet mixing parameter ('elasticNetParam') to find the optimal combination.

This hyperparameter grid, when coupled with a CrossValidator performing 10-fold cross-validation, allows for thorough testing and validation of the model, ensuring robustness and accuracy. Once the cross-validation process is complete, we identify and extract the best-performing LR model. This model is then tested on a separate test dataset to assess its performance, specifically measuring the AUC-ROC to evaluate its effectiveness in sentiment classification. Finally, we report the AUC-ROC score along with the best parameters for regularization and ElasticNet, providing insights into the model's performance and the impact of the chosen hyperparameters.

2. Linear Support Vector Classifier

The Linear Support Vector Classifier (Linear SVC) model in our sentiment analysis project follows a structured and methodical approach. Beginning with defining the Linear SVC model, we set 'word_vectors' as the features and 'sentiment' as the label. The model is then incorporated into a pipeline for streamlined processing. The evaluation of the Linear SVC model is conducted using a Multiclass Classification Evaluator, with accuracy as the chosen metric. To optimize the model's performance, hyperparameter tuning is carried out. A parameter grid is established, varying the regularization parameter ('regParam') and the maximum number of iterations ('maxIter'). This grid includes values of 0.01, 0.1, and 0.2 for 'regParam', and 10, 50, and 100 for 'maxIter', covering a range of possible configurations. After training the model with cross-validation on the training dataset, we identify the best-performing model based on the 10-fold cross-validation results. This model is then tested on a separate test dataset. The primary metric for assessment is accuracy, calculated to determine how well the model performs on unseen data.

3. Random Forest

The architecture of the Random forest Classifier involves creating multiple decision trees during training and outputting the class that is the mode of the classes (classification) of individual trees, thereby improving accuracy and reducing the risk of overfitting. Our implementation of RFC starts with defining the RandomForestClassifier in Apache Spark, with 'word_vectors' as the features and 'sentiment' as the label. This model forms part of a pipeline for systematic processing. For evaluation, we use the

MulticlassClassificationEvaluator, focusing on metrics like accuracy and the F1 score. These metrics are crucial for assessing the model's predictive power and balance between precision and recall. We use a ParamGridBuilder to experiment with different values of 'numTrees' (the number of trees in the forest) and 'maxDepth' (the maximum depth of each tree). Specifically, we explore combinations of 10 and 100 trees and tree depths of 5 and 10. This range allows us to assess the impact of tree complexity and ensemble size on performance. After training with 10-fold cross-validation, we select the best-performing RFC model based on these results. This model is then evaluated on a separate test dataset. The key metrics for assessment are accuracy and the F1 score, providing a holistic view of the model's performance in classifying sentiments.

4. Multi-Layer Perceptron

MLP is a class of feedforward artificial neural networks that consist of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. This structure enables the model to capture complex patterns and relationships in the data. In implementing the MLP Classifier, we first define the architecture of the neural network. Our model consists of three layers: an input layer with 100 nodes (corresponding to the size of our word vectors), a hidden layer with 50 nodes, and an output layer with 2 nodes (for binary classification of sentiment). This setup aims to process the input word vectors through these layers to effectively classify sentiments. e use the ParamGridBuilder to explore different combinations of 'blockSize' (the size of the blocks the input data is divided into during training) and 'maxIter' (the maximum number of iterations). Specifically, we test block sizes of 128 and 256 and maximum iterations of 100 and 200. After training with 10-fold cross-validation, we select the best-performing MLP model based on these results. This model is then evaluated on a separate test dataset. We calculate the accuracy and the F1 score to determine the model's effectiveness in sentiment classification.

Experiments and Results

A. Classification of sentiment

We conducted sentiment classification experiments using three distinct textual sources from IMDb reviews: **the review summary, review detail, and a combination of both summary and detail**. This comprehensive approach aimed to explore the nuances of sentiment analysis from different perspectives and text lengths. We employed four machine learning models—Logistic Regression, Linear SVC, Random Forest Classifier, and Multi-Layer Perceptron—to analyze these text sources. Each model was rigorously trained and tested on each type of text data to evaluate its effectiveness in accurately classifying sentiments, thereby providing a holistic understanding of sentiment trends across different sections of IMDb reviews.

NOTE-(The Plotted Graphs for Hyperparameter Tuning are added in Appendix-End of Report)

1. Logistic Regression

Best Hyperparameter

Case	Parameter	Value
Using Review Summary	**RegParam, ElasticNetParam**	**0.01. 0.0**
Using Review Detail	**RegParam, ElasticNetParam**	**0.01. 0.0**
Using Review Summary and Review Detail	**RegParam, ElasticNetParam**	**0.01. 0.0**

AUC -ROC

Case	Metric	Value
Using Review Summary	**AUC-ROC**	**0.668161**
Using Review Detail	**AUC-ROC**	**0.885838**
Using Review Summary and Review Detail	**AUC-ROC**	**0.892139**

2. Linear SVC

Best Hyperparameter

Case	Parameter	Value
Using Review Summary	**RegParam,MaxIter**	**0.01,50**
Using Review Detail	**RegParam,MaxIter**	**0.01,50**
Using Review Summary and Review Detail	**RegParam,MaxIter**	**0.01,100**

Accuracy

Case	Metric	Value
Using Review Summary	**Accuracy**	**0.737040**
Using Review Detail	**Accuracy**	**0.844862**

Using Review Summary and Review Detail	**Accuracy**	**0.850551**

3. Random Forest Classifier

Best Hyperparameter

Case	Parameter	Value
Using Review Summary	**NumTrees, MaxDepth**	**100,10**
Using Review Detail	**NumTrees, MaxDepth**	**100,10**
Using Review Summary and Review Detail	**NumTrees, MaxDepth**	**100,10**

Accuracy and F1 Score

Case	Metric	Value
Using Review Summary	**Accuracy, F1 Score**	**0.755575, 0.672968**
Using Review Detail	**Accuracy, F1 Score**	**0.811184, 0.7811059**
Using Review Summary and Review Detail	**Accuracy, F1 Score**	**0.8103310, 0.780064**

4. Multi-Layer Perceptron

Best Hyperparameter

Case	Parameter	Value
Using Review Summary	**BlockSize, MaxIter**	**256, 200**
Using Review Detail	**BlockSize, MaxIter**	**128, 100**
Using Review Summary and Review Detail	**BlockSize, MaxIter**	**256, 200**

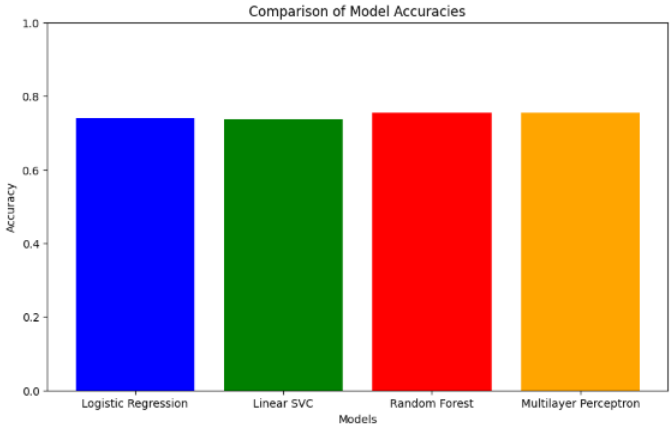
Accuracy and F1 Score

Case	Metric	Value
Using Review Summary	**Accuracy, F1 Score**	**0.754776, 0.6863522**
Using Review Detail	**Accuracy, F1 Score**	**0.8444077, 0.838785**
Using Review Summary and Review Detail	**Accuracy, F1 Score**	**0.857792, 0.845422**

NOTE-The Cross-Validation of the model's Accuracy Graph is added to the Appendix at the end of the report.

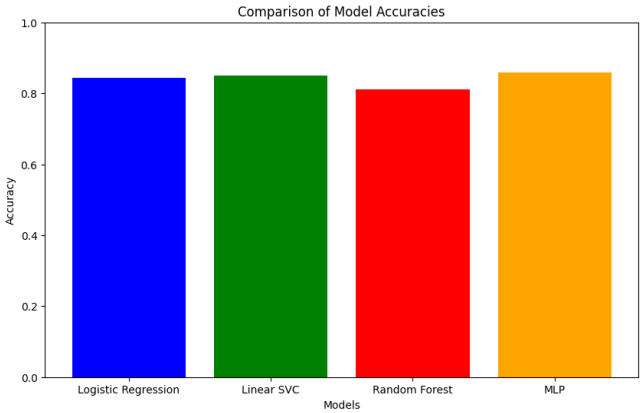
Comparing Models Accuracy across Review Summary

Using Review Summary



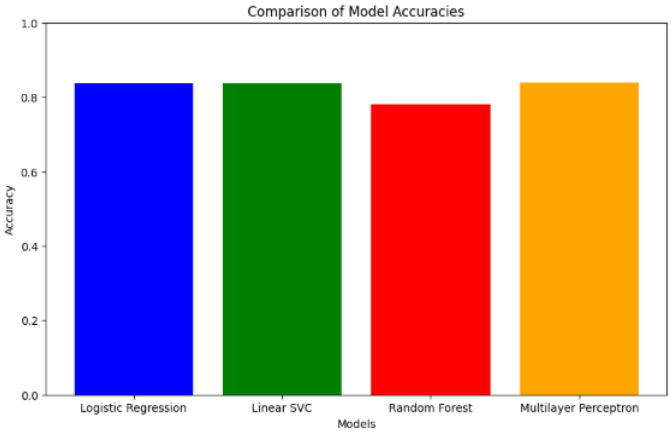
Random Forest has the highest Accuracy

Using Review Detail



Linear SVC and MLP contain almost same Accuracy(Highest)

Using Review Detail and Review Summary(Combined) - MLP has highest Accuracy



B. Identifying Helpful reviewers

We analyzed the dataset to identify the most helpful and least helpful reviewers on IMDb, based on their review upvotes (helpful votes) and downvotes (non-helpful votes) from the **Helpful column** in the dataset. We first transformed the 'helpful' column to separate the **count of upvotes and downvotes** for each review. We then aggregated these votes at the reviewer level to calculate the total helpful and non-helpful votes for each individual. By computing the net helpful votes (helpful minus non-helpful), we were able to gauge the overall reception of each reviewer's contributions. Further, we sorted these reviewers to identify those with the highest positive and negative net votes, providing insights into the most and least appreciated reviewers in the community. Additionally, we calculated the 'helpful_rate', a metric representing the ratio of net helpful votes to the number of reviews written by a reviewer. This rate helped in assessing the overall helpfulness of reviewers on a relative scale. Finally, we sorted reviewers based on this helpful rate to highlight those consistently providing valuable insights (most positive votes rate) and those whose reviews were often found less useful (most negative votes rate). This comprehensive analysis offered a nuanced understanding of reviewer influence and community reception on the IMDb platform.

Reviewers with the best positive record and negative record

Reviewers with the Most Positive Votes Rate

reviewer	total_helpful_votes	total_non_helpful_votes	net_helpful_votes	review_count	helpful_rate
rutan07	951	1	950	1	950.0
s458862411	923	1	922	1	922.0
eak-1	885	1	884	1	884.0
hermes1-2	879	1	878	1	878.0
GVH0	813	1	812	1	812.0
Shut_Up_Irwin	778	1	777	1	777.0
jmkelly03	726	1	725	1	725.0
EvilAdam	715	1	714	1	714.0
mattfg	713	1	712	1	712.0
fernandolindblom	544	56	488	2	244.0
SJ_1	986	105	881	8	110.125

Reviewers with the Most Negative Votes Rate

reviewer	total_helpful_votes	total_non_helpful_votes	net_helpful_votes	review_count	helpful_rate
ccthemovieman-1	9719	13384	-3665	196	-18.698979591836736
MartinHafer	2894	5851	-2957	316	-9.35759493670886
lee_eisenberg	4136	7020	-2884	537	-5.370577281191807
Theo Robertson	2828	5706	-2878	289	-9.95847750865052
jotix100	8085	10750	-2665	317	-8.406940063091483
jboothmillard	722	3273	-2551	357	-7.1456582633053225
bkoganbing	8735	11239	-2504	356	-7.033707865168539
moonspinner55	3472	5855	-2383	313	-7.613418530351438
claudio_carvalho	4460	6647	-2187	250	-8.748
MovieAddict2016	2583	4382	-1799	250	-7.196
noralee	2696	4360	-1664	449	-3.706013363028953
Coventry	2422	4040	-1618	211	-7.668246445497631
movieman_kev	1623	3203	-1580	462	-3.41991341991342
Nazi_Fighter_David	3704	5257	-1553	91	-17.065934065934066
Captain_Couth	2231	3643	-1412	348	-4.057471264367816
The_Void	2245	3643	-1398	224	-6.241071428571429
Spuzzlightyear	1292	2631	-1339	187	-7.160427807486631
rebeljenn	1007	2245	-1238	218	-5.678899082568807
Boba_Fett1138	2367	3515	-1148	186	-6.172043010752688
poolandrews	772	1909	-1137	226	-5.030973451327434

only showing top 20 rows

Conclusion

In conclusion, our project successfully leveraged the rich dataset of IMDb reviews to perform an in-depth sentiment analysis and community feedback evaluation. By employing various machine learning models—Logistic Regression, Linear SVC, Random Forest Classifier, and Multi-Layer Perceptron—we could effectively classify sentiments from different sections of reviews, namely the summary and detail parts. Each model demonstrated its unique strengths and provided valuable insights into how different textual data contribute to sentiment analysis.

We understand that Review Detail provided the models with much better information to train as the accuracies for models with Review Detail had a much higher accuracy compared to Review Summary. This is as expected as the Review Detail is an elaboration about their summaries by the reviewer. The review Summary contained vague descriptions regarding the reviews of the customer and while MLP was able to provide a good accuracy of 75%, it isn't good enough to predict the sentiment of the reviews of customers.

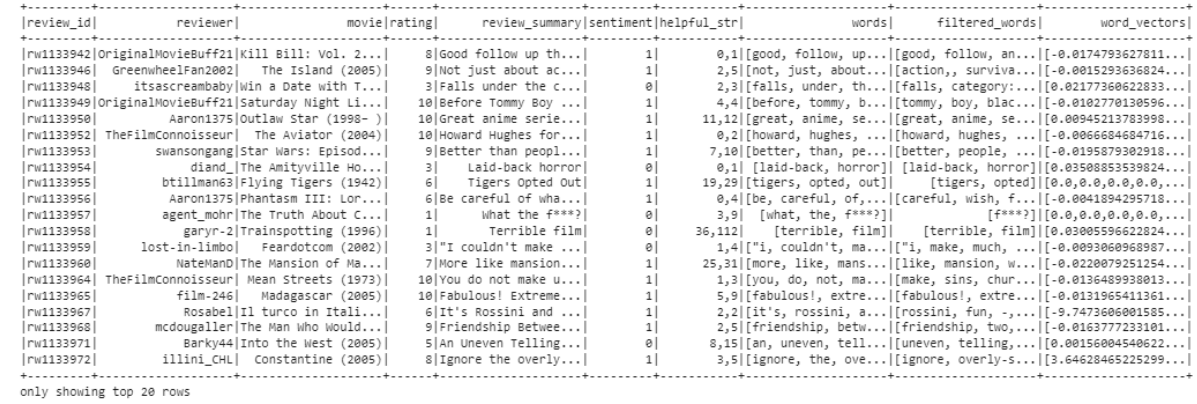
Furthermore, our exploration into the helpfulness of reviewers, using upvotes and downvotes, revealed significant patterns in community engagement and preferences. This aspect of the project not only highlighted the most and least appreciated reviewers but also offered a new perspective on how audience feedback correlates with content quality. The methodologies and findings of this project have broad implications. They can be applied to enhance recommendation systems, improve content curation, and drive a better understanding of audience sentiment in the entertainment industry. This project stands as a testament to the power of data-driven insights in understanding and leveraging user-generated content in digital platforms.

Appendix

All Output Screenshots

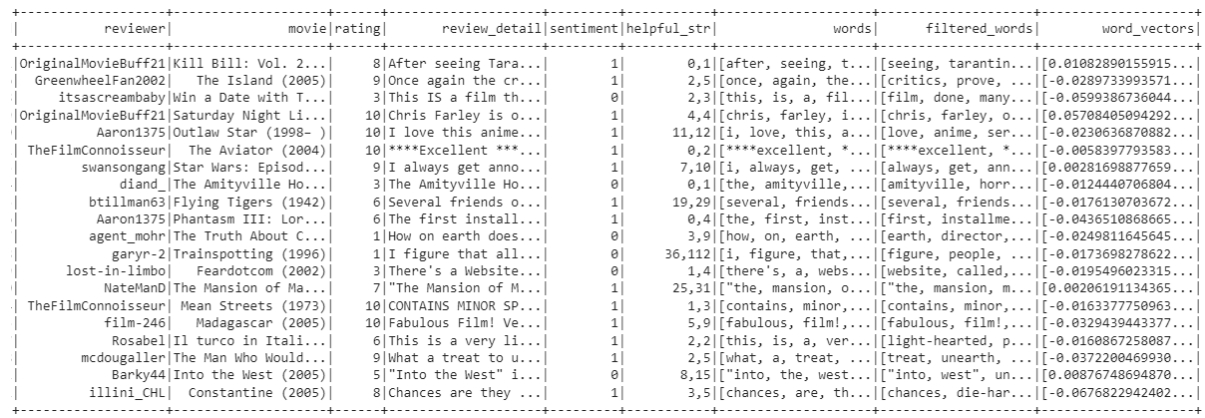
Tokenizing, Encoding(Word2Vector) and removing Stopwords

For Review Summary



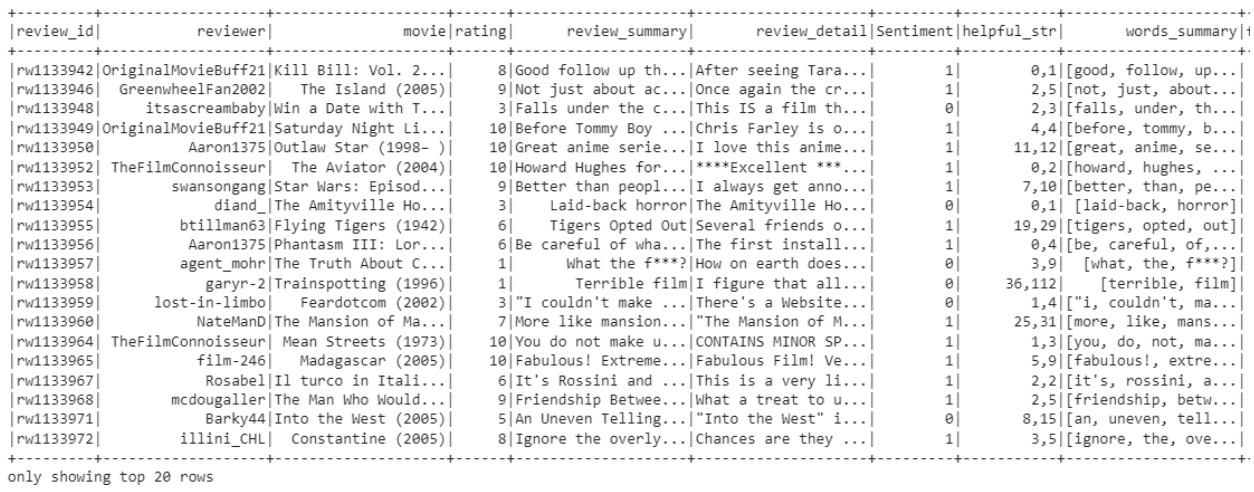
```
+---------+------------------+--------------------+------+--------------+---------+----------+--------------------+--------------------+--------------------+
|review_id|          reviewer|               movie|rating|review_summary|sentiment|helpful_str|               words|      filtered_words|        word_vectors|
+---------+------------------+--------------------+------+--------------+---------+----------+--------------------+--------------------+--------------------+
|rw1133942|OriginalMovieBuff21|Kill Bill: Vol. 2...|     8|Good follow up th...|        1|       0,1|[good, follow, up...|[good, follow, an...|[-0.0174793627811...|
|rw1133946|  GreenwheelFan2002|   The Island (2005)|     9|Not just about ac...|        1|       2,5|[not, just, about...|[action,, surviva...|[-0.0015293636824...|
|rw1133948|     itsascreambaby|Win a Date with T...|     3|Falls under the c...|        0|       2,3|[falls, under, th...|[falls, category:...|[0.02177360622833...|
|rw1133949|OriginalMovieBuff21|Saturday Night Li...|    10|Before Tommy Boy ...|        1|       4,4|[before, tommy, b...|[tommy, boy, blac...|[-0.0102770130596...|
|rw1133950|          Aaron1375|Outlaw Star (1998- )|    10|Great anime serie...|        1|     11,12|[great, anime, se...|[great, anime, se...|[0.00945213783998...|
|rw1133952| TheFilmConnoisseur|  The Aviator (2004)|    10|Howard Hughes for...|        1|       0,2|[howard, hughes, ...|[howard, hughes, ...|[-0.0066684684716...|
|rw1133953|        swansongang|Star Wars: Episod...|     9|Better than peopl...|        1|      7,10|[better, than, pe...|[better, people, ...|[-0.0195879302918...|
|rw1133954|            diand_|The Amityville Ho...|     3|    Laid-back horror|        0|       0,1| [laid-back, horror]| [laid-back, horror]|[0.0350853539824...|
|rw1133955|        btillman63|Flying Tigers (1942)|     6|    Tigers Opted Out|        1|     19,29|[tigers, opted, out]|    [tigers, opted]|[0.0,0.0,0.0,0.0,...|
|rw1133956|         Aaron1375|Phantasm III: Lor...|     6|Be careful of wha...|        1|       0,4|[be, careful, of,...|[careful, wish, f...|[-0.0041894295718...|
|rw1133957|        agent_mohr|The Truth About C...|     1|      What the f***?|        0|       3,9|  [what, the, f***?]|            [f***?]|[0.0,0.0,0.0,0.0,...|
|rw1133958|           garyr-2|Trainspotting (1996)|     1|       Terrible film|        0|    36,112|    [terrible, film]|    [terrible, film]|[0.0300559622824...|
|rw1133959|     lost-in-limbo|   Feardotcom (2002)|     3|"I couldn't make ...|        0|       1,4|["i, couldn't, ma...|["i, make, much, ...|[-0.0093060968987...|
|rw1133960|          NateManD|The Mansion of Ma...|     7|More like mansion...|        1|     25,31|[more, like, mans...|[like, mansion, w...|[-0.0220079251254...|
|rw1133964| TheFilmConnoisseur| Mean Streets (1973)|    10|You do not make u...|        1|       1,3|[you, do, not, ma...|[make, sins, chur...|[-0.0136489938013...|
|rw1133965|          film-246|   Madagascar (2005)|    10|Fabulous! Extreme...|        1|       5,9|[fabulous!, extre...|[fabulous!, extre...|[-0.0131965411361...|
|rw1133967|           Rosabel|Il turco in Itali...|     6|It's Rossini and ...|        1|       2,2|[it's, rossini, a...|[rossini, fun, -,...|[-9.7473606001585...|
|rw1133968|       mcdougaller|The Man Who Would...|     9|Friendship Betwee...|        1|       2,5|[friendship, betw...|[friendship, two,...|[-0.0163777233101...|
|rw1133971|           Barky44|Into the West (2005)|     5|An Uneven Telling...|        0|       8,15|[an, uneven, tell...|[uneven, telling,...|[0.0015600454062...|
|rw1133972|       illini_CHL|  Constantine (2005)|     8|Ignore the overly...|        1|       3,5|[ignore, the, ove...|[ignore, overly-s...|[3.64628465225299...|
+---------+------------------+--------------------+------+--------------+---------+----------+--------------------+--------------------+--------------------+
only showing top 20 rows
```

For Review Detail



```
+------------------+--------------------+------+--------------+---------+----------+--------------------+--------------------+--------------------+
|          reviewer|               movie|rating|review_detail|sentiment|helpful_str|               words|      filtered_words|        word_vectors|
+------------------+--------------------+------+--------------+---------+----------+--------------------+--------------------+--------------------+
|OriginalMovieBuff21|Kill Bill: Vol. 2...|     8|After seeing Tara...|        1|       0,1|[after, seeing, t...|[seeing, tarantin...|[0.01082890155915...|
|  GreenwheelFan2002|   The Island (2005)|     9|Once again the cr...|        1|       2,5|[once, again, the...|[critics, prove, ...|[-0.0289733993571...|
|     itsascreambaby|Win a Date with T...|     3|This IS a film th...|        0|       2,3|[this, is, a, fil...|[film, done, many...|[-0.0599386736044...|
|OriginalMovieBuff21|Saturday Night Li...|    10|Chris Farley is o...|        1|       4,4|[chris, farley, i...|[chris, farley, o...|[0.0570840509429 2...|
|          Aaron1375|Outlaw Star (1998- )|    10|I love this anime...|        1|     11,12|[i, love, this, a...|[love, anime, ser...|[-0.0230636870882...|
| TheFilmConnoisseur|  The Aviator (2004)|    10|****Excellent ***...|        1|       0,2|[****excellent, *...|[****excellent, *...|[-0.0058397793583...|
|        swansongang|Star Wars: Episod...|     9|I always get anno...|        1|      7,10|[i, always, get, ...|[always, get, ann...|[0.00281698877659...|
|            diand_|The Amityville Ho...|     3|The Amityville Ho...|        0|       0,1|[the, amityville,...|[amityville, horr...|[-0.0124440706804...|
|        btillman63|Flying Tigers (1942)|     6|Several friends o...|        1|     19,29|[several, friends...|[several, friends...|[-0.0176130703672...|
|         Aaron1375|Phantasm III: Lor...|     6|The first install...|        1|       0,4|[the, first, inst...|[first, installme...|[-0.0436510868665...|
|        agent_mohr|The Truth About C...|     1|How on earth does...|        0|       3,9|[how, on, earth, ...|[earth, director,...|[-0.0249811645645...|
|           garyr-2|Trainspotting (1996)|     1|I figure that all...|        0|    36,112|[i, figure, that,...|[figure, people, ...|[-0.0173698278622...|
|     lost-in-limbo|   Feardotcom (2002)|     3|There's a Website...|        0|       1,4|[there's, a, webs...|[website, called,...|[-0.0195496023315...|
|          NateManD|The Mansion of Ma...|     7|"The Mansion of M...|        1|     25,31|["the, mansion, o...|["the, mansion, m...|[0.00206191134365...|
| TheFilmConnoisseur| Mean Streets (1973)|    10|CONTAINS MINOR SP...|        1|       1,3|[contains, minor,...|[contains, minor,...|[-0.0163377750963...|
|          film-246|   Madagascar (2005)|    10|Fabulous Film! Ve...|        1|       5,9|[fabulous, film!,...|[fabulous, film!,...|[-0.0329439443377...|
|           Rosabel|Il turco in Itali...|     6|This is a very li...|        1|       2,2|[this, is, a, ver...|[light-hearted, p...|[-0.0160867258087...|
|       mcdougaller|The Man Who Would...|     9|What a treat to u...|        1|       2,5|[what, a, treat, ...|[treat, unearth, ...|[-0.0372200469930...|
|           Barky44|Into the West (2005)|     5|"Into the West" i...|        0|       8,15|["into, the, west...|["into, west", un...|[0.00876748694870...|
|       illini_CHL|  Constantine (2005)|     8|Chances are they ...|        1|       3,5|[chances, are, th...|[chances, die-har...|[-0.0676822942402...|
+------------------+--------------------+------+--------------+---------+----------+--------------------+--------------------+--------------------+
```

For Combined Column Vector of Review Detail and Review Summary

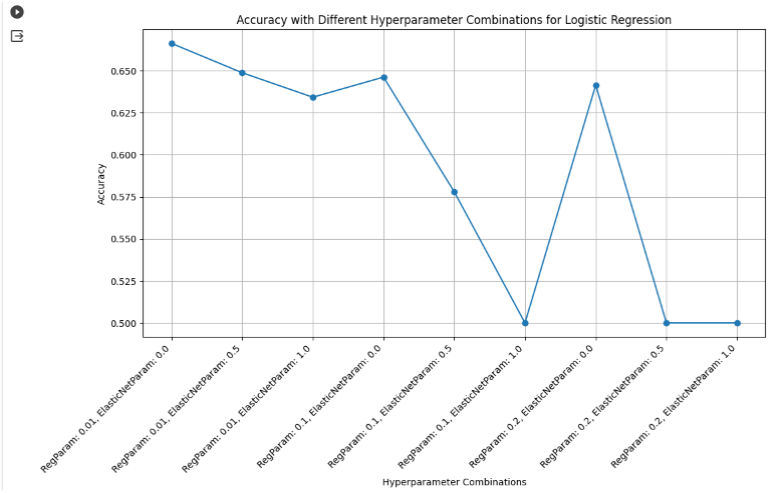


```
+---------+------------------+--------------------+------+--------------+--------------+---------+----------+--------------------+--------------------+
|review_id|          reviewer|               movie|rating|review_summary|review_detail|Sentiment|helpful_str|       words_summary|                |
+---------+------------------+--------------------+------+--------------+--------------+---------+----------+--------------------+--------------------+
|rw1133942|OriginalMovieBuff21|Kill Bill: Vol. 2...|     8|Good follow up th...|After seeing Tara...|        1|       0,1|[good, follow, up...|
|rw1133946|  GreenwheelFan2002|   The Island (2005)|     9|Not just about ac...|Once again the cr...|        1|       2,5|[not, just, about...|
|rw1133948|     itsascreambaby|Win a Date with T...|     3|Falls under the c...|This IS a film th...|        0|       2,3|[falls, under, th...|
|rw1133949|OriginalMovieBuff21|Saturday Night Li...|    10|Before Tommy Boy ...|Chris Farley is o...|        1|       4,4|[before, tommy, b...|
|rw1133950|          Aaron1375|Outlaw Star (1998- )|    10|Great anime serie...|I love this anime...|        1|     11,12|[great, anime, se...|
|rw1133952| TheFilmConnoisseur|  The Aviator (2004)|    10|Howard Hughes for...|****Excellent ***...|        1|       0,2|[howard, hughes, ...|
|rw1133953|        swansongang|Star Wars: Episod...|     9|Better than peopl...|I always get anno...|        1|      7,10|[better, than, pe...|
|rw1133954|            diand_|The Amityville Ho...|     3|    Laid-back horror|The Amityville Ho...|        0|       0,1| [laid-back, horror]|
|rw1133955|        btillman63|Flying Tigers (1942)|     6|    Tigers Opted Out|Several friends o...|        1|     19,29|[tigers, opted, out]|
|rw1133956|         Aaron1375|Phantasm III: Lor...|     6|Be careful of wha...|The first install...|        1|       0,4|[be, careful, of,...|
|rw1133957|        agent_mohr|The Truth About C...|     1|      What the f***?|How on earth does...|        0|       3,9|  [what, the, f***?]|
|rw1133958|           garyr-2|Trainspotting (1996)|     1|       Terrible film|I figure that all...|        0|    36,112|    [terrible, film]|
|rw1133959|     lost-in-limbo|   Feardotcom (2002)|     3|"I couldn't make ...|There's a Website...|        0|       1,4|["i, couldn't, ma...|
|rw1133960|          NateManD|The Mansion of Ma...|     7|More like mansion...|"The Mansion of M...|        1|     25,31|[more, like, mans...|
|rw1133964| TheFilmConnoisseur| Mean Streets (1973)|    10|You do not make u...|CONTAINS MINOR SP...|        1|       1,3|[you, do, not, ma...|
|rw1133965|          film-246|   Madagascar (2005)|    10|Fabulous! Extreme...|Fabulous Film! Ve...|        1|       5,9|[fabulous!, extre...|
|rw1133967|           Rosabel|Il turco in Itali...|     6|It's Rossini and ...|This is a very li...|        1|       2,2|[it's, rossini, a...|
|rw1133968|       mcdougaller|The Man Who Would...|     9|Friendship Betwee...|What a treat to u...|        1|       2,5|[friendship, betw...|
|rw1133971|           Barky44|Into the West (2005)|     5|An Uneven Telling...|"Into the West" i...|        0|      8,15|[an, uneven, tell...|
|rw1133972|       illini_CHL|  Constantine (2005)|     8|Ignore the overly...|Chances are they ...|        1|       3,5|[ignore, the, ove...|
+---------+------------------+--------------------+------+--------------+--------------+---------+----------+--------------------+--------------------+
only showing top 20 rows
```

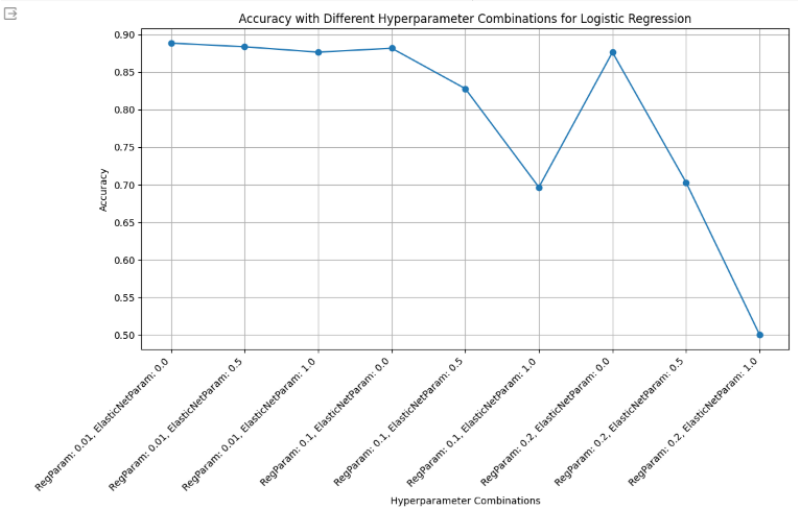
words_summary	filtered_words_summary	word_vectors_summary	words_detail	filtered_words_detail	word_vectors_detail	combined_features
[good, follow, up...]	[good, follow, an...]	[0.09897127184376...]	[after, seeing, t...]	[seeing, tarantin...]	[-0.0534763250230...]	[0.09897127184376...]
[not, just, about...]	[action,, surviva...]	[0.01554072275757...]	[once, again, the...]	[critics, prove, ...]	[-0.0394247354402...]	[0.01554072275757...]
[falls, under, th...]	[falls, category:...]	[0.00530465831980...]	[this, is, a, fil...]	[film, done, many...]	[-0.0367161307949...]	[0.00530465831980...]
[before, tommy, b...]	[tommy, boy, blac...]	[0.02498977037224...]	[chris, farley, i...]	[chris, farley, o...]	[-0.0049703205869...]	[0.02498977037224...]
[great, anime, se...]	[great, anime, se...]	[0.08315026491181...]	[i, love, this, a...]	[love, anime, ser...]	[-0.0695174744544...]	[0.08315026491181...]
[howard, hughes, ...]	[howard, hughes, ...]	[-0.0010966970415...]	[****excellent, *...]	[****excellent, *...]	[0.01360663533210...]	[-0.0010966970415...]
[better, than, pe...]	[better, people, ...]	[0.01190652512013...]	[i, always, get, ...]	[always, get, ann...]	[-0.0375068674133...]	[0.01190652512013...]
[laid-back, horror]	[laid-back, horror]	[0.05178220197558...]	[the, amityville,...]	[amityville, horr...]	[-0.0457603408650...]	[0.05178220197558...]
[tigers, opted, out]	[tigers, opted]	[0.0,0.0,0.0,0.0,...]	[several, friends...]	[several, friends...]	[-0.0621036605940...]	(200,[100,101,102...
[be, careful, of,...]	[careful, wish, f...]	[-0.0010488545522...]	[the, first, inst...]	[first, installme...]	[-0.0476437096521...]	[-0.0010488545522...]
[what, the, f***?]	[f***?]	[0.0,0.0,0.0,0.0,...]	[how, on, earth, ...]	[earth, director,...]	[-0.0183711961554...]	(200,[100,101,102...
[terrible, film]	[terrible, film]	[-0.0592315085232...]	[i, figure, that,...]	[figure, people, ...]	[-0.0770213876745...]	[-0.0592315085232...]
["i, couldn't, ma...]	["i, make, much, ...]	[0.01091368713726...]	[there's, a, webs...]	[website, called,...]	[-0.0369525389039...]	[0.01091368713726...]
[more, like, mans...]	[like, mansion, w...]	[6.45740399098334...]	["the, mansion, o...]	["the, mansion, m...]	[-0.0591593121164...]	[6.45740399098334...]
[you, do, not, ma...]	[make, sins, chur...]	[0.02045345166698...]	[contains, minor,...]	[contains, minor,...]	[-0.0019397507119...]	[0.02045345166698...]
[fabulous!, extre...]	[fabulous!, extre...]	[-0.0049627900356...]	[fabulous, film!,...]	[fabulous, film!,...]	[-0.0691767041713...]	[-0.0049627900356...]
[it's, rossini, a...]	[rossini, fun, -,...]	[0.0170787090912...]	[this, is, a, ver...]	[light-hearted, p...]	[-0.0157078844057...]	[0.0170787090912...]
[friendship, betw...]	[friendship, two,...]	[-0.0125896851532...]	[what, a, treat, ...]	[treat, unearth, ...]	[-0.0137132407894...]	[-0.0125896851532...]
[an, uneven, tell...]	[uneven, telling,...]	[0.02426180820912...]	["into, the, west...]	["into, west", un...]	[-0.0425362927424...]	[0.02426180820912...]
[ignore, the, ove...]	[ignore, overly-s...]	[7.19500162328283...]	[chances, are, th...]	[chances, die-har...]	[0.00543170812797...]	[7.19500162328283...]

Building a Logistic Regression

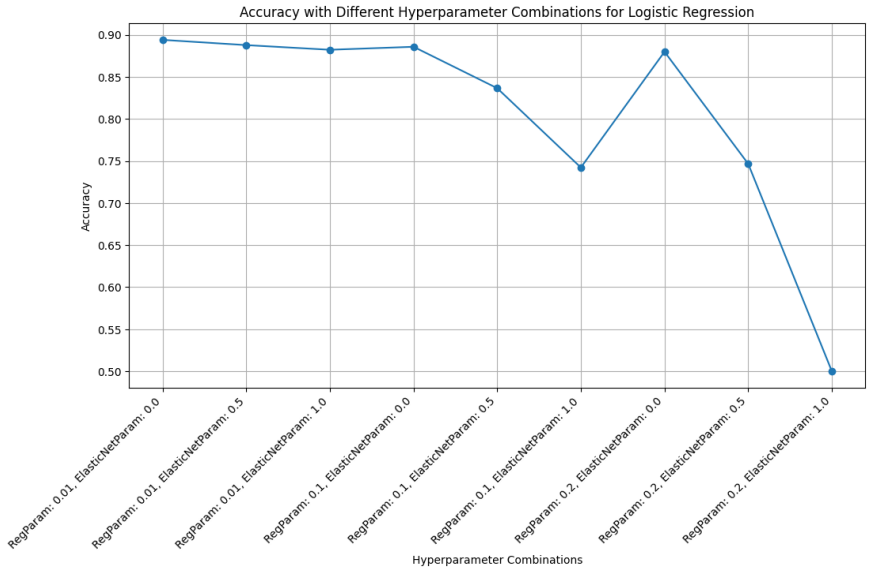
Using Review Summary



Using Review Detail

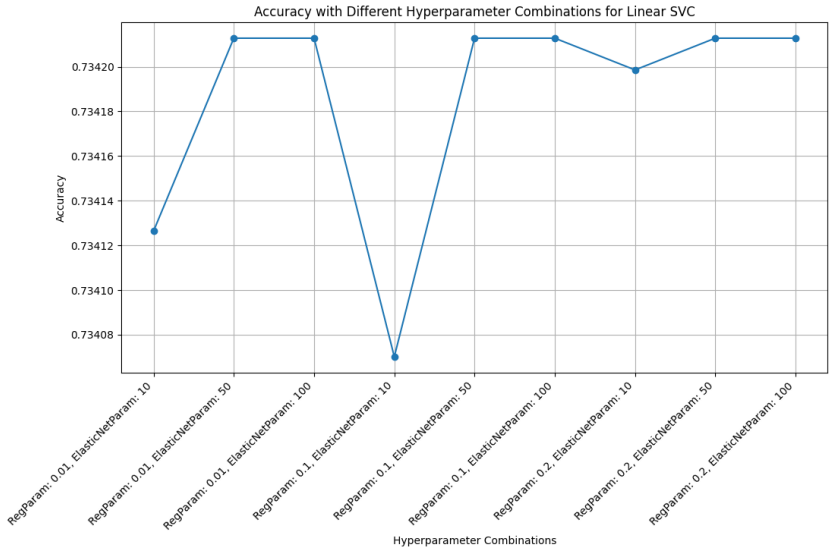


Using Review Detail and Review Summary

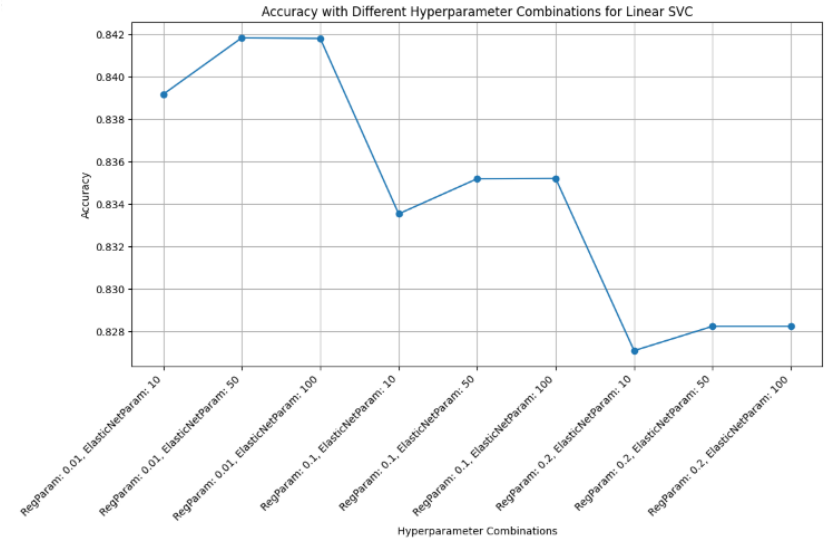


Building a Linear SVC

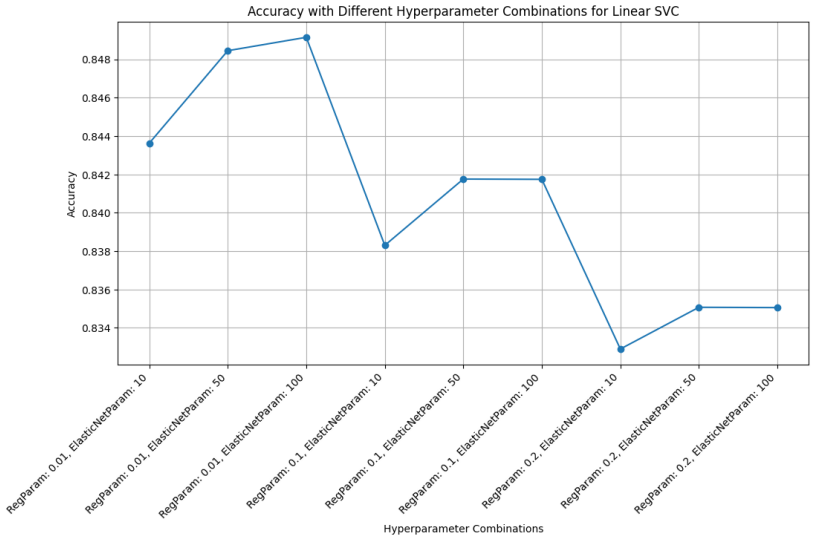
Using Review Summary



Using Review Detail

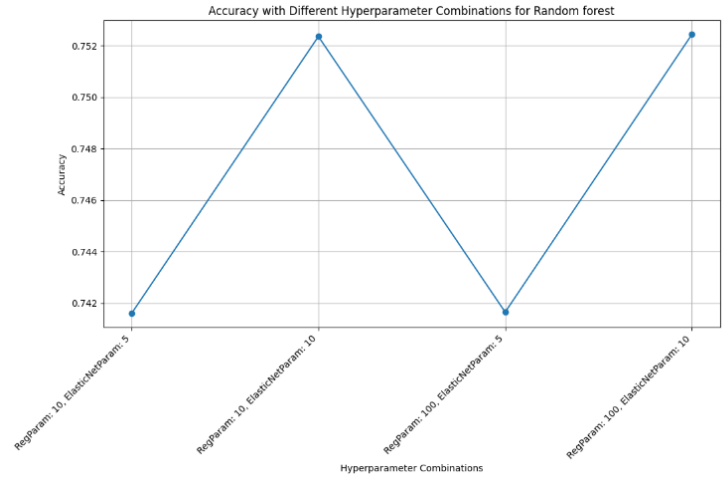


Using Review Detail and Review Summary

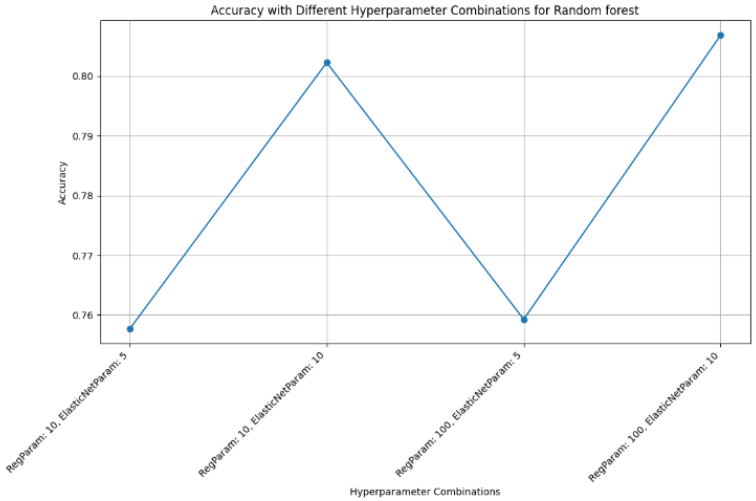


Building a Random forest

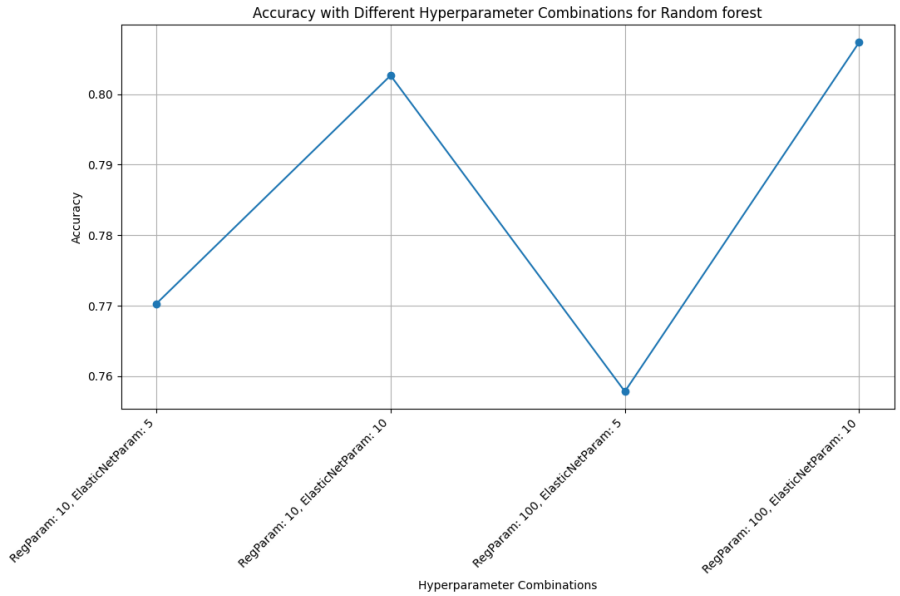
Using Review Summary



Using Review Detail

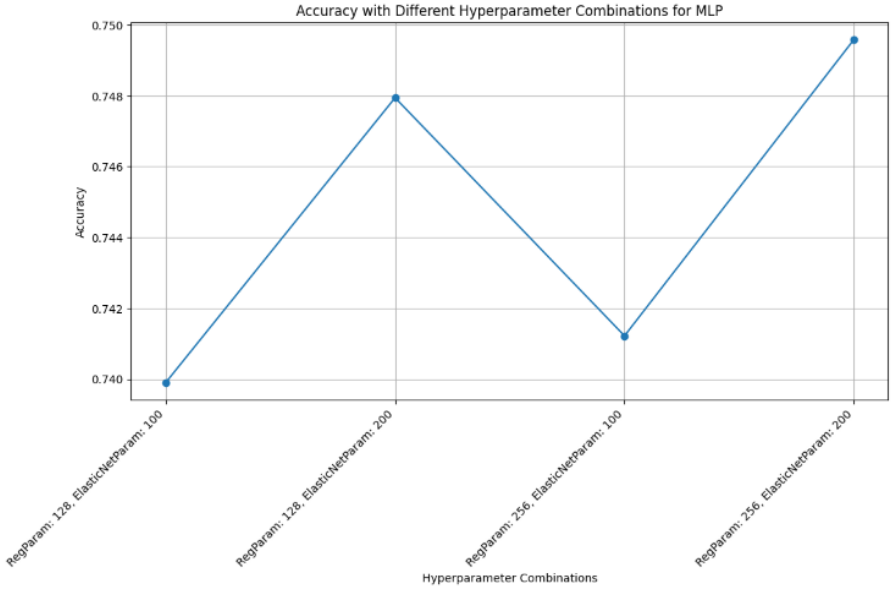


Using Review Detail and Review Summary

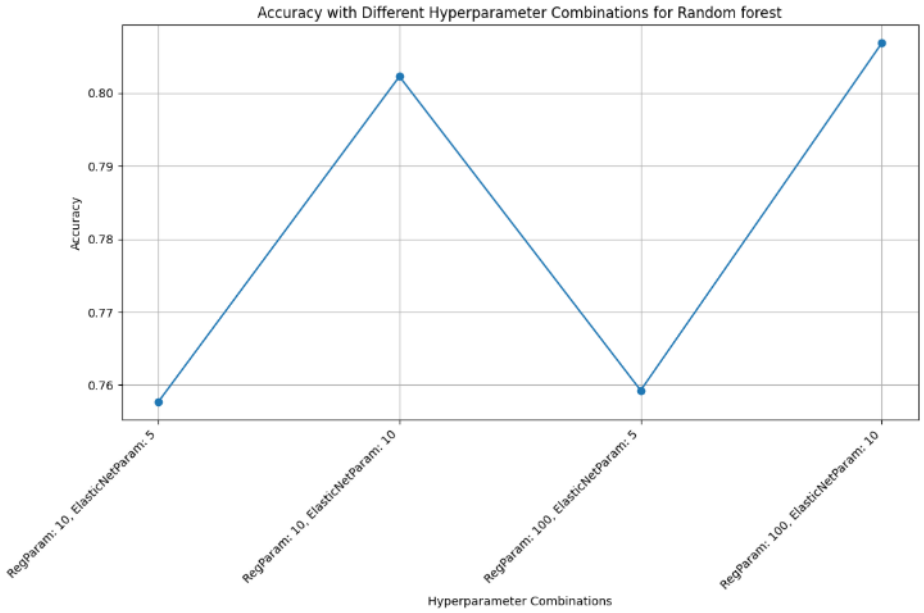


Building a Multi-Layer Perceptron

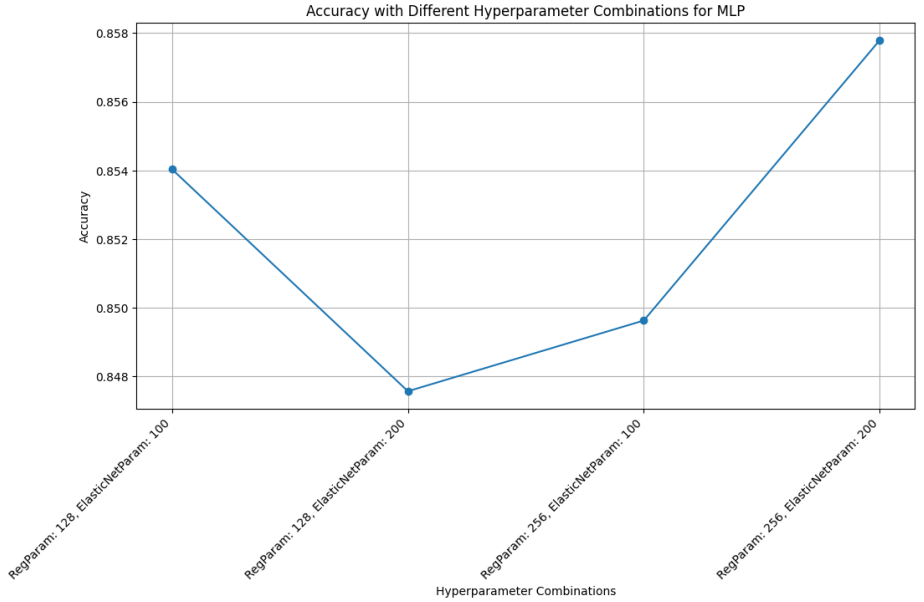
Using Review Summary



Using Review Detail



Using Review Detail and Review Summary



HELPFUL VOTES RECEIVED BY THE REVIEWER

Total Helpful votes and Not Helpful Votes received

reviewer	total_helpful_votes	total_non_helpful_votes	net_helpful_votes
sweetmik	83	163	-80
edcrosay	0	0	0
janmarc-1	0	9	-9
milesjgraham	19	40	-21
awnorm	2	11	-9
tetrical	2	9	-7
rmax304823	650	1082	-432
CherryBlossomBoy	2	5	-3
jhclues	15	37	-22
lizzie_mag	22	44	-22
imonroe	11	24	-13
jrsmitty	3	7	-4
Glamdring-The-Foe...	2	3	-1
coyets	0	0	0
turgaytemel	0	1	-1
danielri	0	0	0
eyzarblu-1	1	6	-5
fwendt	4	9	-5
ales-mrak	4	16	-12
electromance	0	1	-1

only showing top 20 rows

Reviewers with the most Net Positive and Net Negative votes

Reviewers with the Most Positive Votes

reviewer	total_helpful_votes	total_non_helpful_votes	net_helpful_votes
rutan07	951	1	950
s458862411	923	1	922
eak-1	885	1	884
SJ_1	986	105	881
hermes1-2	879	1	878
GVH0	813	1	812
Shut_Up_Irwin	778	1	777
jmkelly03	726	1	725
EvilAdam	715	1	714
mattfg	713	1	712
fernandolindblom	544	56	488

Reviewers with the Most Negative Votes

reviewer	total_helpful_votes	total_non_helpful_votes	net_helpful_votes
ccthemovieman-1	9719	13384	-3665
MartinHafer	2894	5851	-2957
lee_eisenberg	4136	7020	-2884
Theo Robertson	2828	5706	-2878
jotix100	8085	10750	-2665
jboothmillard	722	3273	-2551
bkoganbing	8735	11239	-2504
moonspinner55	3472	5855	-2383
claudio_carvalho	4460	6647	-2187
MovieAddict2016	2583	4382	-1799
noralee	2696	4360	-1664
Coventry	2422	4040	-1618
movieman_kev	1623	3203	-1580
Nazi_Fighter_David	3704	5257	-1553
Captain_Couth	2231	3643	-1412
The_Void	2245	3643	-1398
Spuzzlightyear	1292	2631	-1339
rebeljenn	1007	2245	-1238
Boba_Fett1138	2367	3515	-1148
poolandrews	772	1909	-1137

only showing top 20 rows

Rate of the Positive and negative votes received

reviewer	total_helpful_votes	total_non_helpful_votes	net_helpful_votes	review_count	helpful_rate
sweetmik	83	163	-80	1	-80.0
edcrosay	0	0	0	1	0.0
janmarc-1	0	9	-9	1	-9.0
milesjgraham	19	40	-21	1	-21.0
awnorm	2	11	-9	1	-9.0
tetrical	2	9	-7	1	-7.0
rmax304823	650	1082	-432	75	-5.76
CherryBlossomBoy	2	5	-3	3	-1.0
jhclues	15	37	-22	2	-11.0
lizzie_mag	22	44	-22	1	-22.0
imonroe	11	24	-13	2	-6.5
jrsmitty	3	7	-4	2	-2.0
Glamdring-The-Foe...	2	3	-1	1	-1.0
coyets	0	0	0	1	0.0
turgaytemel	0	1	-1	1	-1.0
danielri	0	0	0	1	0.0
eyzarblu-1	1	6	-5	1	-5.0
fwendt	4	9	-5	1	-5.0
ales-mrak	4	16	-12	6	-2.0
electromance	0	1	-1	1	-1.0

only showing top 20 rows