Employing Text Mining for distinguishing American and Non-American movie categories

Juliana Gouveia de Sá Couto
TDDE16 | Text Mining
Linköping University
julgo420@student.liu.se

Abstract

This project aims to use machine learning to sort movies into American and non-American categories, aiming to uncover cultural storytelling patterns in their plots. Through extensive exploratory data analysis and the application of various models, including logistic regression, support vector machine, and random forest, the analysis achieves notable success. By extensively analyzing the data and applying various models like logistic regression, support vector machine, and random forest, the project succeeds notably. Specifically, the logistic regression model, considering word combinations of 1-3 grams, achieves an impressive F1-micro score of 0.864. This outperforms the basic models, indicating better precision and recall. The project demonstrates that machine learning can effectively distinguish cultural differences in movie narratives. The success of the logistic regression model, especially in analyzing word combinations, highlights its ability to understand and classify cultural patterns in diverse storytelling styles. This achievement holds promise for improving our grasp of cultural nuances in movies and shows the potential of machine learning in recognizing and categorizing these patterns.

1 Introduction

This project addresses the task of classifying movies into American and non-American categories using diverse machine learning models. The focal point is decoding subtle cultural nuances within movie plots, aiming to identify distinctive storytelling patterns indicative of cultural origins. Beyond a conventional classification challenge, the project delves into cultural analysis through artificial intelligence, seeking to automate the recognition of cultural storytelling cues.

The problem's intrinsic interest lies in its potential to unveil cultural identifiers within the cinematic landscape. By automating the recognition of these narrative elements, the project goes beyond categorization, providing a perspective on cultural analysis. The automated identification of cultural storytelling cues holds promise for platforms aiming to enhance user experiences through personalized and culturally relevant content recommendations.

Solving this problem provides insights into the feasibility of leveraging machine learning for cultural analysis. The successful classification of movies based on cultural patterns demonstrates the adaptability of these models to discern intricate linguistic and narrative features. This project emphasizes the fusion of technology and culture, showcasing machines' potential not just for categorization but for the interpretation of the complex narratives that shape our global cultures.

Can machine learning models discern cultural storytelling patterns within movie plots, and if so, to what extent?

2 Theory

The study of cultural storytelling patterns within movie plots involves a multidisciplinary approach, integrating concepts from natural language processing (NLP), machine learning, and data analysis. While some concepts align with standard techniques covered in introductory courses, this section delves into advanced methods and theoretical frameworks that provide a deeper understanding of the key processes employed in this project.

2.1 Text Mining and Natural Language Processing (NLP)

Text mining and NLP form the bedrock of this analysis, facilitating the extraction of meaningful insights from unstructured text data. The use of spaCy and NLTK libraries enables advanced text processing, including tokenization, lemmatization, and collocation identification.

The application of collocations enhances the accuracy of subsequent analyses. Collocations capture nuances in cultural storytelling by identifying word pairs that convey specific cultural references or themes within movie plots. To unearth these meaningful word associations, a scoring threshold is established based on Mutual Information (MI) scores. The BigramCollocationFinder from the NLTK library is employed to sift through the movie plots, pinpointing word pairs that exhibit significant co-occurrence. By choosing a scoring threshold, in this case, a MI score of 4, the preprocessing pipeline ensure that only substantial and relevant collocations are retained, filtering out noise and irrelevant word pairs.

Once identified, these collocations are replaced within the movie plots to create a modified and enriched dataset. This replacement involves merging the collocated words into a single entity, often connected by an underscore. This transformation aims at preserving the integrity of the collocations while simplifying the dataset's structure for more effective analysis.

2.2 Topic Modeling with Latent Dirichlet Allocation (LDA)

LDA is a probabilistic generative model used for discovering hidden thematic structures within a collection of documents. In the context of analyzing movie plots, LDA enables the identification of latent topics, representing recurring themes or narrative elements. LDA operates on the assumption that each document is a mixture of a small number of topics, and each word in the document is attributable to one of those topics. This probabilistic model allows for a more nuanced exploration of the underlying themes in movie plots.

Coherence metrics are crucial for determining the optimal number of topics generated by LDA. UMass and C_V are measures of the interpretability and meaningfulness of topics. UMass evaluates the degree of semantic similarity between words within the same topic, while C_V assesses the coherence by considering the relative distances between words. By optimizing these metrics, the analysis ensures that the identified topics are not only numerically optimal but also semantically meaningful, providing a foundation for a richer understanding of cultural storytelling nuances.

2.3 Document Embedding with Doc2Vec

Doc2Vec captures the context and semantics of words in a document, offering an innovative approach to embedding documents into high-dimensional vectors. This method preserves contextual information, addressing the limitations of traditional bag-of-words models when dealing with cultural storytelling. Doc2Vec's consideration of word order and relationships makes it well-suited for capturing the intricate storytelling styles that vary across cultures.

2.4 Dimensionality Reduction and Visualization

PCA is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional representation while retaining as much of the original variance as possible. In the analysis of Doc2Vec embeddings, PCA is employed to breakdown the essential information and reveal underlying patterns in the relationships between movie plots, particularly in terms of duration and origin.

The introduction of Bootstrap Sampling Significance Tests underscores the commitment to a comprehensive exploration of diverse models. This approach goes beyond a simple performance comparison by providing a statistically grounded understanding of whether observed differences in model performance are likely to be meaningful or could occur by chance.

2.5 Most Frequent Label Heuristic

The classification task starts with the Most Frequent Label Heuristic, a straightforward baseline assigning the most common label in the training set to all instances in the test set. This heuristic serves as a benchmark for evaluating the relative improvements achieved by more advanced models in discerning cultural nuances within movie plots.

2.6 Bootstrap Sampling Significance Tests

To gauge the statistical significance of observed differences in model performance, the analysis turns to Bootstrap Sampling Significance Tests. This technique, grounded in resampling theory, involves creating multiple subsamples with replacement from the original dataset. The theoretical underpinning lies in estimating the distribution of a statistic by drawing random samples. By comparing observed differences to those that could occur by chance, the analysis seeks to determine the statistical significance of disparities in model performance, providing insights into the robustness of the cultural pattern recognition achieved by the diverse set of machine learning models.

This approach goes beyond a simple performance comparison by providing a statistically grounded understanding of whether observed differences in model performance are likely to be meaningful or could occur by chance.

3 Data

3.1 Data Source

In this project, the primary dataset utilized is the *'wiki_movie_plots_deduped.csv'* obtained from Kaggle's "Wikipedia Movie Plots" dataset. This dataset contains a lot of information, including plot summaries from Wikipedia for approximately 34,886 movies on a global scale. Each entry in the dataset encompasses essential details such as the release year, movie title, origin/ethnicity, director(s), main actors and actresses, movie genre(s), Wikipedia page URL, and a detailed plot description. Here's an example:

- Release year: 2006;
- Movie title: Casino Royale;
- Origin/ethnicity: American;
- Director(s): Martin Campbell;
- Main actors and actresses: Daniel Craig, Eva Green, Mads Mikkelsen, Judi Dench, Giancarlo Giannini, Jeffrey Wright;
- Movie genre(s): Action;
- Wikipedia page URL: https://en.wikipedia.org/wiki/Casino_Royale_(2006_film)
- Detailed plot description: "MI6 agent James Bond gains his licence to kill and status as a 00 agent by assassinating the traitorous MI6 section chief Dryden at the British Embassy in Prague, as well as his terrorist contact, Fisher, in a bathroom in Lahore, Pakistan. (...)"

(Note: the transcription in the last bullet point corresponds to the first sentence of the description in the "Plot" section of the Wikipedia page. No more has been transcribed due to its length)

The data source provides a diverse and comprehensive collection of movie plots from 1910 to 2019, allowing for a broad exploration of cinematic content across different genres, cultural origins, and time periods. It is important to note that this richness in information is fundamental for conducting analyses related to content-based movie recommendation, text classification, and other natural language processing applications.

3.2 Data Creation Process

In addition to the data already present in the previously mentioned csv, the final dataset used in this project was obtained through a meticulous web scrapping process, aiming to the extraction of crucial information about film running times from Wikipedia.

The procedure involved accessing the HTML content of each Wikipedia page and subsequently analysing it using BeautifulSoup. Specifically, the script identified the 'Running time' information within the infobox of the Wikipedia page. This information was then compiled and stored in a new column labeled 'durations' in the dataset.

To refine the dataset and start processing the data, observations that did not contain duration information were systematically eliminated. In addition, to quantify execution times in a standardised way, numerical representations of durations in minutes were extracted using regular expressions. This additional information was incorporated into a new column labelled "minutes", enriching the dataset with a crucial quantitative measure.

Here's an example from the 1903 film "The Great Train Robbery", which on the Wikipedia page has the following running time: 740 ft (230 m) 12 minutes (at 18 frame/s). It was then converted to 12 minutes.

3.3 Preprocessing Strategies and Rationale

In the data preprocessing phase, a sequence of transformative steps was employed to refine the raw movie plot data, ensuring its suitability for subsequent analyses. These steps aimed at standardizing the representation of words, eliminating noise, and enhancing the meaningfulness of the dataset. By systematically addressing issues such as case sensitivity, numerical variations, and irrelevant symbols, the preprocessing laid the foundation for a more robust and coherent analysis of movie plots.

- **Lowercasing**: The text was normalized to lowercase, ensuring uniformity in word representations throughout the dataset. This step prevents the model from treating words with different cases as distinct.

- **Number Removal**: All numerical characters were removed from the movie plot text. This step helps eliminate numerical variations, ensuring that the model focuses on the semantic meaning of the words rather than numerical values.

- **Punctuation Removal**: Punctuation marks were eliminated from the text to avoid irrelevant variations and maintain consistency in word representation. This step aids in accurate analysis by reducing noise in the data.

- **Stopword Removal**: Common stopwords (e.g., 'the', 'and', 'is') were removed to exclude frequently occurring but less meaningful words. This enhances the relevance of words in the dataset for subsequent analysis.

- **Lemmatization**: Words in the text were lemmatized, reducing them to their base or root form. This step minimizes variability in word forms, providing a more coherent and standardized dataset for analysis.

- **Collocation Identification and Replacement**: Meaningful word combinations (collocations) in movie plots were identified and replaced. This enhances the accuracy of subsequent analyses by capturing and preserving significant word associations in the text.

- **Proper Noun Removal**: Proper nouns were identified and removed using POS tagging. This step is crucial for mitigating bias in topic modeling, ensuring that proper names do not disproportionately influence the analysis, and preserving the indicative nature of the movie plots.

Here, as an example, is the plot of the film Casino Royale after pre-processing: agent james bond gain licence kill status agent assassinate traitorous section chief dryden british embassy prague terrorist contact fisher bathroom lahore pakistan.

To end the data preparation phase, two crucial binary target variables, namely *'american'* and *'is_long'*, were meticulously constructed to lay the foundation for subsequent classification tasks. The *'american'* variable categorizes movies based on their origin, distinguishing between American and non-American productions. Simultaneously, *'is_long'* classifies movies as either long or short based on their duration, with a threshold set at 100 minutes. These variables serve as pivotal attributes for the predictive models that will follow, adding a layer of classification to the dataset.

Following this, the dataset undergoes a strategic partitioning into training, development, and test sets, each constituting 70%, 15%, and 15% of the total data, respectively. These datasets ensure the robustness of subsequent analyses, model training and prediction, laying the groundwork for a comprehensive exploration of movie plot patterns.

3.4 Statistics after preprocessing

The preprocessing of document data has significantly reduced the dataset's size and complexity. Initially, there were 29,580 documents with 11,556,993 tokens and 370,627 unique types (unique words or terms present in the dataset). After tasks like tokenization, stemming, and stopword removal, the dataset still has 29,580 documents, but with a reduced token count of 5,526,867 and only 110,798 unique types.

This reduction in tokens and types indicates that preprocessing has streamlined the dataset by removing unnecessary information, standardizing text representations, and improving efficiency for tasks like topic modeling and document embedding.

The decrease in the number of unique types shows that similar terms have been consolidated, making the dataset more manageable and focused. These statistics highlight the importance and impact of preprocessing in preparing text data for subsequent modeling or analytical tasks.

4 Method

4.1 Stage 1: Exploratory Data Analysis

The research methodology for identifying cultural storytelling patterns in movie plots involved a comprehensive approach to understanding and analyzing the dataset. The exploration started with detailed Exploratory Data Analysis (EDA), examining global movie release trends, analyzing movie duration statistics, and exploring the distribution of movie origins.

To enhance clarity and facilitate examination, similar ethnicities were strategically consolidated into broader country categories by adding a new column to the DataFrame using Pandas.

The analysis then focused on the historical trajectories of leading countries in movie production, specifically the United States and India. Python libraries like Pandas and Matplotlib were used for data manipulation and visualization, extracting insights into industry growth patterns, considering key historical events and shifts. The exploration concluded with an assessment of the prevalence of American and non-American movies, highlighting their distinct thematic preferences through a topic modeling approach.

The topic modeling process involved meticulous tokenization, bigram identification using the Gensim library, and the application of Latent Dirichlet Allocation (LDA) to identify themes distinguishing American and non-American films. Iterative evaluation, considering coherence metrics like UMass and C_V, guided the selection of the optimal number of topics.

Training the LDA model and extracting topic descriptors provided nuanced insights into prevalent themes characterizing American and non-American films, refining the understanding of thematic compositions and uncovering distinct preferences.

The methodology then transitioned to document embedding, specifically using the Doc2Vec model, to evaluate the resulting document matrix for a classification task. However, the visualization of the reduced matrix through Principal Component Analysis (PCA) did not reveal distinct clusters based on movie duration or origin. As a result, the decision was made to prioritize the TFIDF matrix for subsequent stages of the analysis.

4.2 Stage 2: Classification Task

The study aimed to distinguish between American and non-American movies using various machine learning models that focus on narrative patterns indicative of cultural origins. Python libraries like Scikit-learn, Matplotlib, Seaborn, and TensorFlow were used for data manipulation, visualization, and model implementation.

The Most Frequent Label Baseline started with DummyClassifier to predict movie categories based on the most frequent class in the training data. This established a baseline performance level.

The Logistic Regression Baseline involved constructing a TF-IDF matrix with 2-6 gram characters and applying logistic regression models. Parameter tuning explored regularization techniques ('elasticnet,' 'l1,' 'l2,' and 'none') using Scikit-learn's LogisticRegression with the 'saga' solver to improve classification performance.

Another logistic regression model focused on 1-3 grams extracted from movie plots, using Scikit-learn's TF-IDF vectorizer and LogisticRegression. Grid search cross-validation optimized penalties (L1 or L2 regularization) and tuned the penalty parameter (C) for enhanced classification results.

The Support Vector Machine (SVM) section used Scikit-learn's SVC to classify movies. Various kernel types, cost parameters (C), and gamma parameters were tested to identify the optimal SVM configuration.

The Random Forest model employed Scikit-learn's RandomForestClassifier with 'gini' and 'entropy' criteria. This ensemble learning technique combined multiple decision trees' predictions through a majority vote, enhancing robustness. Evaluation metrics, especially the F1-micro score, provided a comprehensive assessment.

In the evaluation phase, detailed confusion matrices were used to examine each baseline and model. These matrices offered a granular visualization of classification outcomes, providing insights into strengths and limitations.

Bootstrap Sampling Significance Tests were integrated for a robust comparative analysis. This statistical technique involved resampling from the original dataset to compute and compare performance scores across diverse subsamples, determining the statistical significance of observed variations in model performance.

Combining confusion matrices and significance testing enriched the overall analysis, contributing to a holistic interpretation of the models' efficacy in distinguishing cultural storytelling patterns in movie plots.

5 Result

5.1 EDA: Movie Production Trends Over the Century

5.1.1 Overall Trend in Movie Releases

The analysis begins with an exploration of the overall trend in movie production over the years. The distribution plot of release years (Figure 1) indicates a significant increase in movie production from the early 20th century, reaching its peak around the mid-20th century. The dataset shows a reduction in movie releases during the 1960s, which could be attributed to the introduction of television. However, a further recovery occurs in the latter part of the century, corresponding wi the beginning of the digital age.

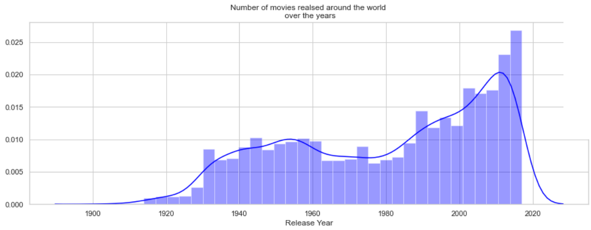

Figure 1: Number of Movies Released Over the Years

5.1.2 Origin Distribution

The exploration of the movie dataset reveals a notable dominance of American movies, with a diverse range of ethnicities and origins represented. That said, among the 24 origins documented in the dataframe, the United States emerges as the unequivocal leader in movie production.

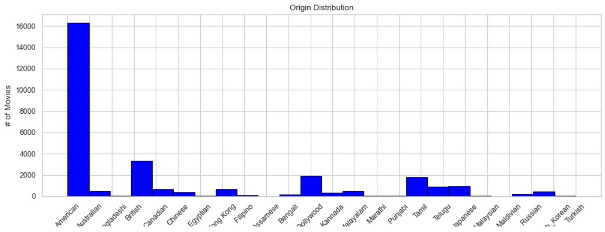

Figure 2: Distribution of Movie Origins

A noteworthy observation is the substantial presence of ethnicities from India, prompting a pragmatic step forward in simplifying the analysis. To facilitate further examination and comprehension, a new column has been added to the dataframe, explicitly denoting the countries corresponding to these diverse ethnicities. This strategic augmentation lays the groundwork for a more streamlined and insightful analysis of the dataset.

5.1.3 Country-wise Movie Production

The analysis delves into the top countries contributing to movie production. The bar chart (Figure 3) illustrates the number of movies produced by different countries, showcasing the leading positions of the United States, India, the United Kingdom, and Japan. The categorical classification of countries into broader categories enhances clarity, revealing that India secures the second position in movie production.

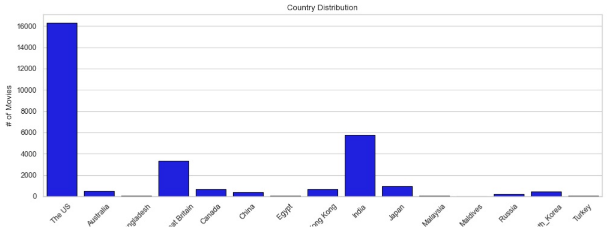

Figure 3: Country-wise Movie Production

To visualize the periods of intensive movie production for different countries, a heatmap (Figure 4) is generated by grouping the data by "Country" and "Release Year." The heatmap confirms the dominance of American movies throughout the years, underlining the United States as a leader in the global movie industry.

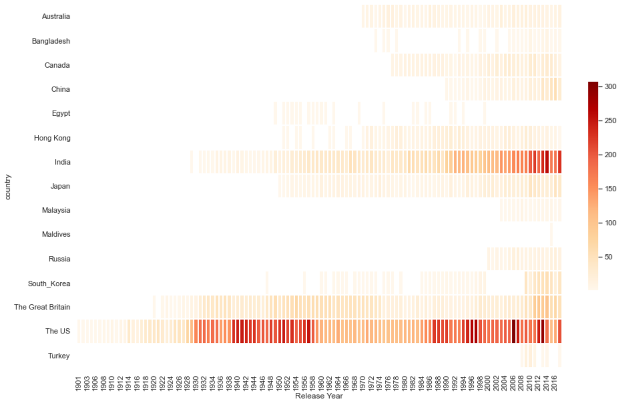

Figure 4: Movie Production Trends by Country and Year

Examining the distribution of movie origins (Figure 5) highlights the dominance of American movies in the dataset. The histogram shows the clear majority of movies originating from the United States, indicating its significant role in the global movie industry. Despite the presence of diverse ethnicities and origins, the United States emerges as the unequivocal leader in movie production among the documented origins.

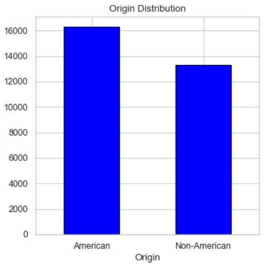

Figure 5: Distribution of Movie Origins

Early cinematic efforts by leading countries like the United States, India, the United Kingdom, and

Japan created significant production variations. In the U.S., the movie industry grew until the 1960s, then declined for two decades due to television and changing audience demographics, leading to the shift from "Old Hollywood" to "New Hollywood." The subsequent recovery was linked to the onset of the digital age, enabling faster and more cost-effective production.

In India, there was steady growth until the 1980s, interrupted by rising violence, declining musical quality, and increased video piracy. The turning point came with the release of "Chandni" in 1989, breaking the trend of violent action films and reviving the romantic musical genre. This shift, termed "New Bollywood," set a new standard. From the 2000s, the Indian film industry saw rapid growth, possibly influenced by India's expanding population and the transformative impact of the digital age on production efficiency.

5.2 EDA: Topic Modeling Analysis

Through iterative evaluation of different topic numbers, the script employs two coherence metrics, UMass and C_V, to gauge the quality of topics. The visualization of coherence scores (Figure 6) guides the selection of the optimal number of topics (4) by striking a balance between UMass and C_V scores, a decision supported by correlation analysis - the higher the scores, the more coherent and interpretable the topics are.

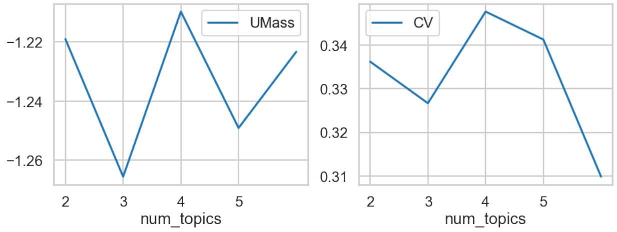

Figure 6: UMass and C_V Coherence Scores

Topic modeling is employed to explore thematic preferences in American and non-American movies. The bar chart (Figure 7) displays the distribution of topics, revealing distinct thematic preferences. Analyzing the distribution of topics revealed that American movies are prominently associated with themes related to crime (topic 4) and action (topic 2). Conversely, non-American films tend to focus more on themes of love and family (topic 3). The first topic appears to be linked to sports and nations, with minimal differentiation between the two classes. That is, American movies are associated with crime and action, while non-American films focus more on themes of love and family.

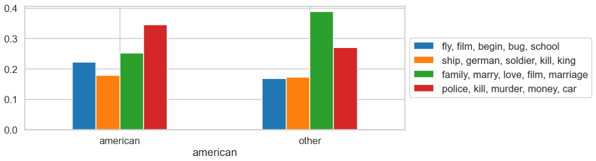

Figure 7: Distribution of Topics in American and Non-American Movies

5.3 EDA: Doc2Vec Model and Dimensionality Reduction PCA Visualization

The Doc2Vec model is trained on a corpus of movie plots, and the resulting document matrix is subjected to dimensionality reduction using Principal Component Analysis (PCA). However, the visualization of the reduced matrix does not reveal distinct clusters based on movie duration or origin. Consequently, the decision is made to not employing the use of the Doc2Vec matrix for the classification task and instead prioritize the TFIDF matrix in the subsequent stages of the analysis.

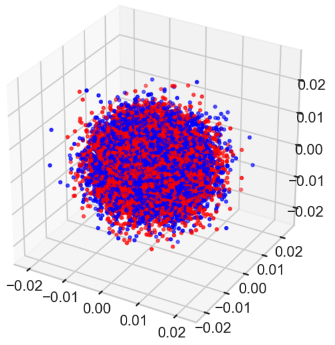

Figure 8: PCA Visualization of Doc2Vec Model

5.4 Classification task: American VS non-American movies

The first baseline F1 score (Most Frequent Label - 0.544) indicates a moderate level of effectiveness for a simple prediction model. The second baseline F1 score (2-6grams Logistic Regression - 0.858) for the Logistic Regression model with specific character combinations performs significantly better, showcasing the model's ability to capture patterns compared to a basic predictor.

After tweaking the Logistic Regression model, the F1 score increased by a small margin (0.006), but it's considered statistically insignificant. The model with 1-3 gram characters achieved an F1 score of 0.862, slightly better than the second baseline, suggesting improved performance with different data configurations.

The Support Vector Machine (SVM) with specific settings achieved the highest F1-micro score (0.860), outperforming other models and demonstrating superior predictive accuracy.

On the other hand, the Random Forest model underperformed compared to other models, suggesting challenges in capturing dataset patterns. Possible reasons include sensitivity to noise, overfitting, or insufficient hyperparameter tuning.

Model	**F1-micro**
Logistic 1-3 grams	0.864773
SVM (rbf, C = 100)	0.860266
Logistic 2-6 char grams (no reg.)	0.858012
Logistic 2-6 char grams	0.855308
Random Forest	0.809105
Most Frequent Label Predictor	0.543160

Table 1: Performance of Different Models on the Test Set

5.4.1 Bootstrap sampling significance tests

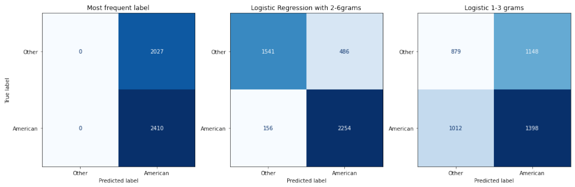

Figure 9: Confusion Matrix - baselines and Logistic Regression

The baseline Logistic Regression model (2-6grams) under performs the 1-3grams Logistic Regression model in terms of F1-micro score on the test set. This significant difference implies that, at least within the scope of this analysis, the Logistic Regression model is a more effective choice for the given classification task than the baseline as the obtained p-value of 0.0 indicates that the observed difference in F1-micro scores between the two models is statistically significant.

The statistical analysis (p-value of 0.0) reveals a significant difference in favor of the 1-3grams Logistic Regression predictor over the Most Frequent Label model in terms of F1-micro score on the test set (Figure 10). The extremely low p-value suggests that the observed discrepancy in performance is highly unlikely to occur by random chance.

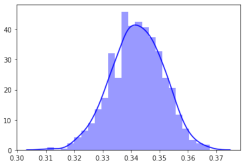

Figure 10: Bootstrap sampling significance tests - Most Frequent Label VS Logistic Regression

6 Discussion

The classification models exhibited varying degrees of success in distinguishing between American and non-American movies based on cultural storytelling patterns. Precision, recall, F1-score, and accuracy metrics were calculated for each model, providing a comprehensive evaluation of their performance. Notably, 1-3grams Logistic Regression demonstrated the highest accuracy, indicating its effectiveness in capturing the nuances of cultural narratives. Key findings from exploratory data analysis, topic modeling, and document embedding were instrumental in informing the interpretation of these results. For instance, the identification of prevalent themes through topic modeling contributed to understanding the thematic compositions that drove classification outcomes.

Despite the promising results, several limitations need consideration. The dataset's potential biases, stemming from the source and selection criteria, could influence the models' generalizability. Assumptions made during preprocessing, such as the removal of proper nouns, might impact the recognition of culturally significant names. Additionally, constraints in the machine learning models, such as the sensitivity to imbalanced datasets, could affect overall performance. It is crucial to acknowledge these limitations to provide a nuanced interpretation of the study's outcomes.

7 Conclusion

In summary, this project aimed to classify movies as American or non-American based on cultural storytelling patterns, using various machine learning models. The approach involved exploring the data, transforming it, conducting historical analysis, topic modeling, document embedding, and a detailed classification task.

The results show that the models effectively identified cultural storytelling patterns in movie plots, achieving high accuracy in distinguishing between American and non-American movies. The models' interpretability, rooted in exploratory data analysis and topic modeling, provided insights into the key features influencing classification.

Key findings highlight the intricate connection between cultural nuances and narrative elements in movie plots. The success of machine learning models in this context demonstrates AI's flexibility in understanding complex linguistic and thematic aspects of cultural storytelling.

Moreover, the project suggests broader applications for automated cultural analysis, beyond categorization, such as content recommendation and customized narrative development. By automating the recognition of cultural storytelling cues, the research showcases how technology can interpret diverse narratives in our global cinematic landscape.

In conclusion, this project not only solved the research problem but also paved the way for further exploration at the intersection of artificial intelligence and cultural analysis. It underscores technology's transformative potential in deciphering and understanding the richness of cultural narratives embedded in movie plots.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Leo Breiman. 2001. Random forests. *Machine Learning*.

Christopher J.C. Burges. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*.

Adele Cutler and Leo Breiman. 2004. Random forests for classification in ecology. *Ecology*.

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

Stefan Evert. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, IMS, University of Stuttgart.

Phillip I. Good. 2005. *Resampling Methods: A Practical Guide to Data Analysis*. Birkhäuser.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *Introduction to Statistical Learning*. Springer.

Ian T. Jolliffe. 2002. *Principal Component Analysis*. Wiley StatsRef: Statistics Reference Online.

Jey Han Lau, David Newman, and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 78–86.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196.

Andreas C. Müller and Sarah Guido. 2016. *An Introduction to Machine Learning with scikit-learn*. O'Reilly Media.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108.

J. Robischon. Unknown. Wikipedia movie plots.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2005. *Introduction to Data Mining*. Addison-Wesley.

Wikipedia. Unknown. The great train robbery (1903 film).