**OPIM 5671: Text Mining (NLP)**

**Medium Blog Post Category Classification**

**Lakshmi Ravi Chandu Kolusu**

**Ms Business Analytics & Project Management**

**Specialization: Data Science**

**Introduction:**

In the dynamic landscape of online content creation and dissemination, Medium stands out as a versatile platform that fosters a diverse array of ideas and perspectives. As a hub for writers spanning various domains, Medium provides an open space for users to share their knowledge, insights, and engage in discussions through published articles. However, as the volume of content on Medium continues to grow, the need for efficient content categorization becomes increasingly crucial.

The focus of this report is on the development of a solution aimed at automating the categorization of blog content on Medium. Specifically, our goal is to leverage textual features present in the titles and subtitles of articles to automatically assign them to relevant categories. The target variable, 'category,' serves as a representation of the general topic of a Medium blog post. By automating the content categorization process, our objective is to enhance the efficiency of the platform, enabling users to more easily discover and engage with content that aligns with their interests.

To accomplish this task, we utilize the Medium Post Titles dataset sourced from Kaggle, comprising Medium post titles and associated categories. This dataset serves as the foundation for training and evaluating our categorization model. Through the application of machine learning and natural language processing techniques, we aim to develop a robust solution that optimizes the user experience on Medium by providing more streamlined access to content across diverse topics.

We aim to automate content categorization, in order to not only be directed at enhancing the user experience on Medium but also contribute to the broader discourse on the intersection of technology and content creation. This report outlines our methodology, challenges faced, and the insights gained throughout the development process of the solution, ultimately highlighting the potential impact on the efficiency and accessibility of online platforms fostering knowledge sharing and collaboration.

**Data Description:**

**Raw Data:**

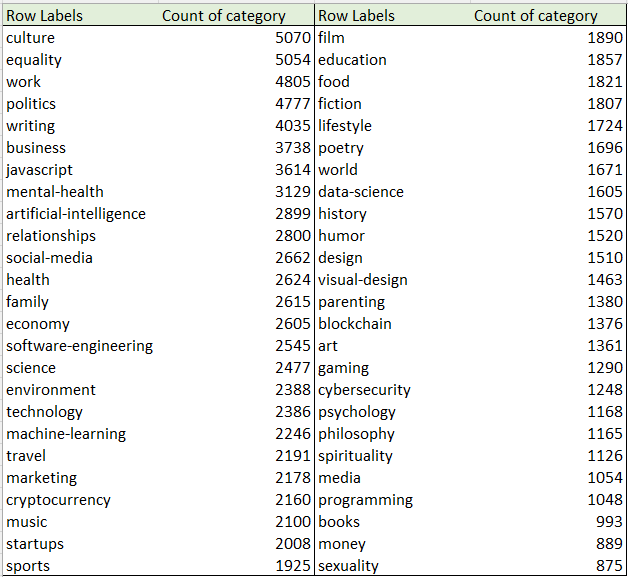
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Source** | **Reference Link** | **Rows** | **Columns** |
| Medium Post Titles | Kaggle | <https://www.kaggle.com/datasets/nulldata/medium-post-titles> | 126,418 | 4 |

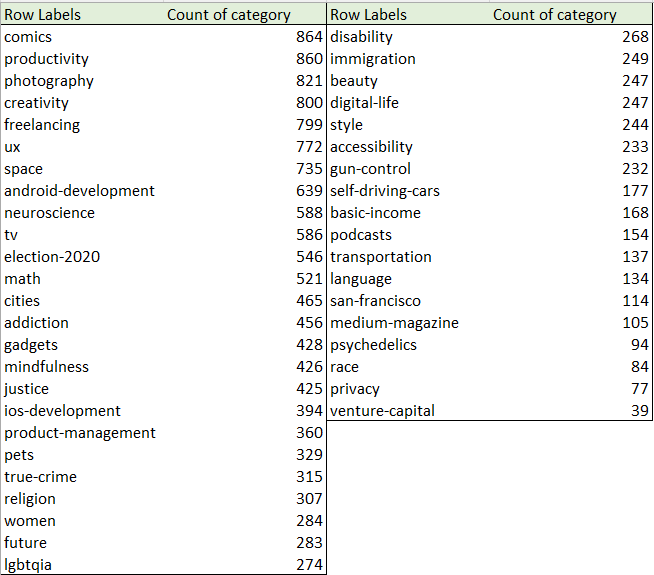
**Overview of Variables:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Variable Name** | **Description** | **Type** |
| Target Variable | Category | Category of the Medium Blog Post | Str |
| Input Variable | Title | Title of the Medium Blog Post | Str |
| Input Variable | Subtitle | Description of the Medium Blog Post | Str |
| Input Variable | Subtitle Truncated Flag | True or False: If the Subtitle has been Truncated | Str |
| Input Variable | Title/Subtitle | Combined Title and Subtitle of the Medium Blog Post | Str |

**Data Exploration & Manipulation:**

The dataset comprises a total of 93 categories, with some categories containing thousands of data points while others have only hundreds. To ensure an adequate volume of data for training while also considering and enhancing the precision of our model, we selected 12 categories to focus on for our model: culture, politics, business, health, economy, science, environment, travel, music, sports, film and food. We considered two key factors in our selection criteria: (1) categories with more than a thousand rows, and (2) the level of generality, comprehensiveness, or distinctiveness of each category. Specifically, categories that are too narrow or specific in scope, such as 'election-2020,' 'blockchain,' 'productivity,' 'creativity,' 'freelancing,' 'android-development,' 'ios-development,' and 'san-francisco,' will be omitted from our final dataset. This process of data selection and filtering is essential to ensure that the model has enough representative data for accurate classification. Throughout the iterations of the model, we reduced the number of categories selected. For the fourth and fifth iteration, the 5 categories chosen are business, culture, economy, health and politics. For the sixth iteration, the 5 categories chosen were politics, economy, science, music and sports.





**Model Building:**

* Model Selection Criteria:

In selecting the most suitable model for automating content categorization on *Medium*, we prioritize robust performance metrics to ensure the efficacy of our solution. Two key criteria guide our model selection process: the Receiver Operating Characteristic (ROC) of the test data and the misclassification rate. The ROC curve provides a comprehensive evaluation of the model's ability to discriminate between different categories, considering both sensitivity and specificity. A higher ROC score indicates superior model performance. Simultaneously, the misclassification rate serves as a crucial measure, highlighting the percentage of incorrectly categorized articles. By emphasizing these metrics, we aim to strike a balance between accurate categorization and minimizing false positives or negatives, ultimately delivering a model that optimally enhances the user experience on Medium.

* Model Refinement:

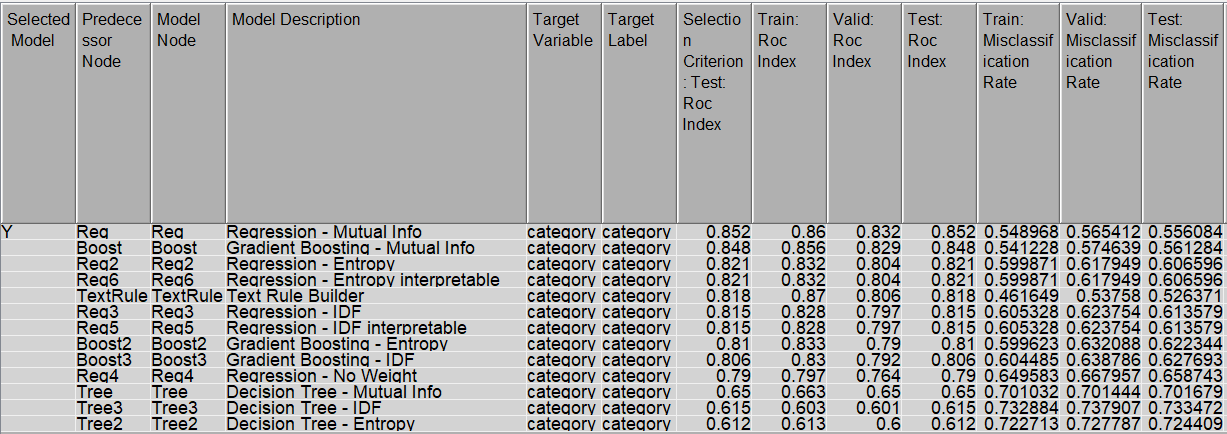
The process of model refinement played a pivotal role in enhancing the efficiency and precision of our text analysis framework. Stemming was employed to streamline words into their root form, promoting the consolidation of related terms. For instance, words like "blog" and "blogs" were reduced to "blog." This not only contributed to a more compact representation but also facilitated a clearer understanding of the underlying content. Additionally, synonym handling was implemented to treat equivalent terms, such as "tv" and "television," as synonyms, thereby further reducing the multitude of unique terms and promoting a more cohesive analysis.

Furthermore, we undertook the task of filtering both common and rare terms to mitigate noise in the dataset. Utilizing a default SAS stop list and the head and tail of the Zipf plot, we eliminated very common words that might lack significant meaning and extremely rare words that could introduce unnecessary complexity. Concept linking was integrated into the refinement process to evaluate associations between terms and filter out those with weak connections to the main categories. Notably, 86 terms were identified and subsequently added to the stop list through adjustments in the keep column of the term filter viewer in the text filter node. This meticulous manual adjustment of topic term cutoff values, coupled with the exploration of various data partition ratios, played a critical role in fine-tuning the model for optimal performance. Additionally, adjustments to text cluster resolution further honed the accuracy and relevance of our text analysis framework. The iterative nature of these refinements reflects our commitment to precision and effectiveness in extracting meaningful insights from textual data.

**Model Selection Process by Iteration & Model Comparison:**

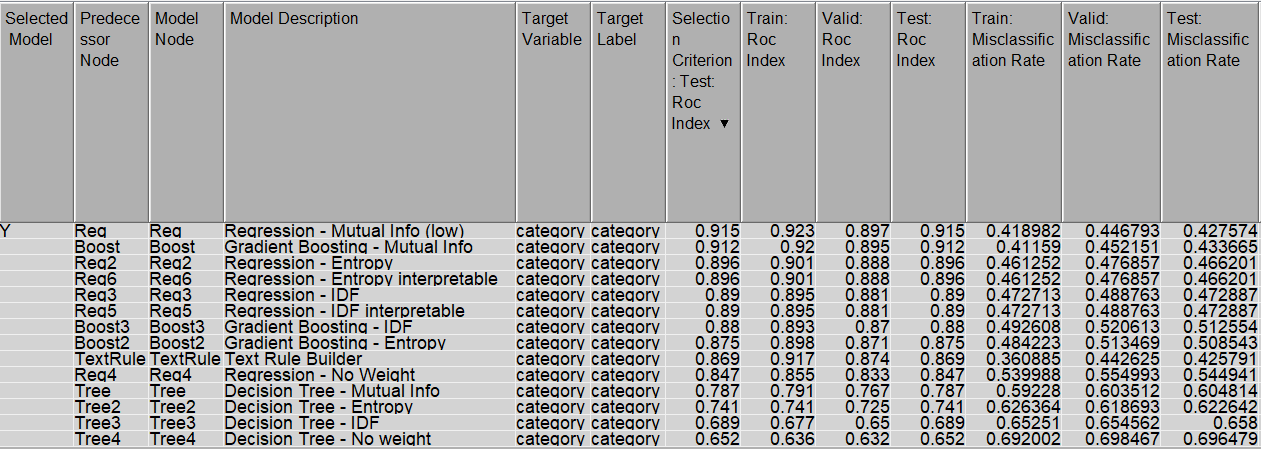
When building our model, we went through 6 iterations to test which would lead to the best ROC and Misclassification Rate.

The first model was a regression model with term weights set at mutual information. This first model is a text classification based on the input variable ‘subtitle.’ The test set provided a ROC of 0.852 and a Misclassification Rate of 0.556.



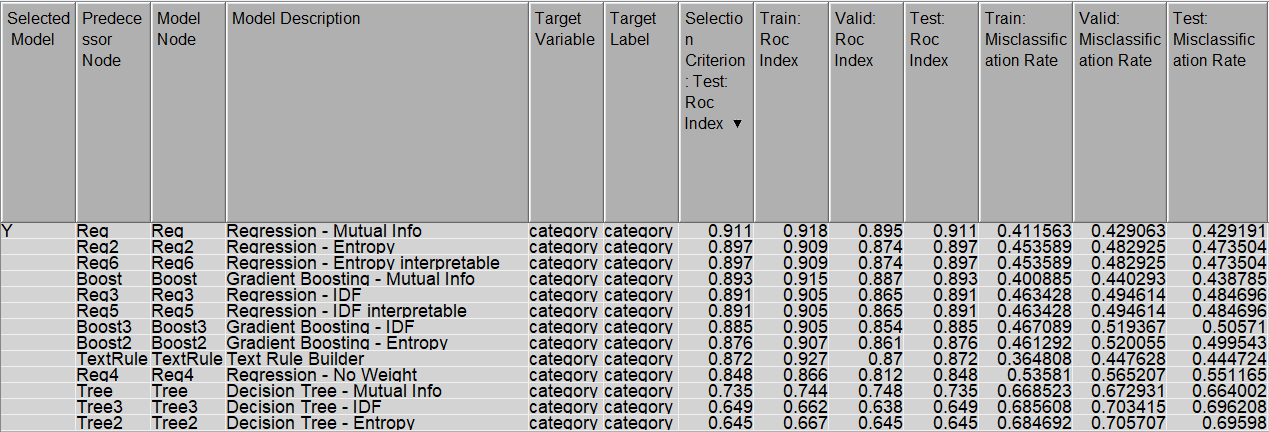
**Table 1: The First Iteration Results**

The second model we tested was also a regression model with term weights set at mutual information, but it is a text classification based on the two input variables of ‘title’ and ‘subtitle’ combined. We thought that by combining these two text fields, we would see greater accuracy from the model. The second iteration did show better performance with a ROC of 0.915 and a Misclassification Rate of 0.428.



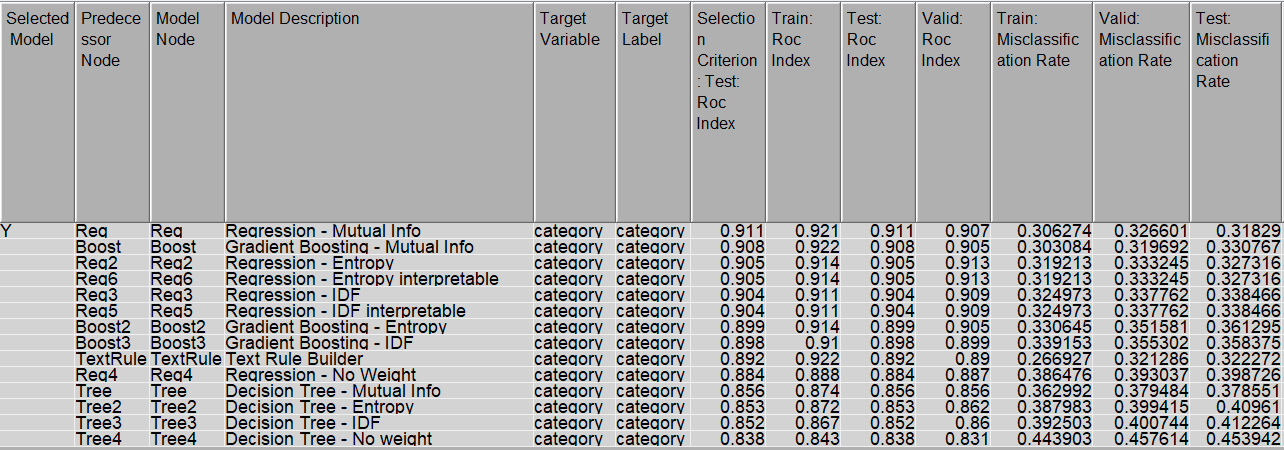
**Table 2: The Second Iteration Results**

For the third iteration, we maintained the model used in the second iteration but introduced undersampling to address potential class imbalance effects on performance. Specifically, we equalized the number of data points across classes by undersampling the majority classes—culture, politics, business, health, economy, science, environment, travel, music, sports, and film—to match the quantity of the minority class, food. The rationale behind this was to assess if a balanced dataset could optimize our model's results. However, the outcomes were not as anticipated. This model exhibited a slightly lower ROC of 0.911 compared to the second iteration and presented a higher Misclassification Rate of 0.429. Despite our attempt to rectify class imbalances, it appears that this approach did not lead to an improvement in overall model performance.



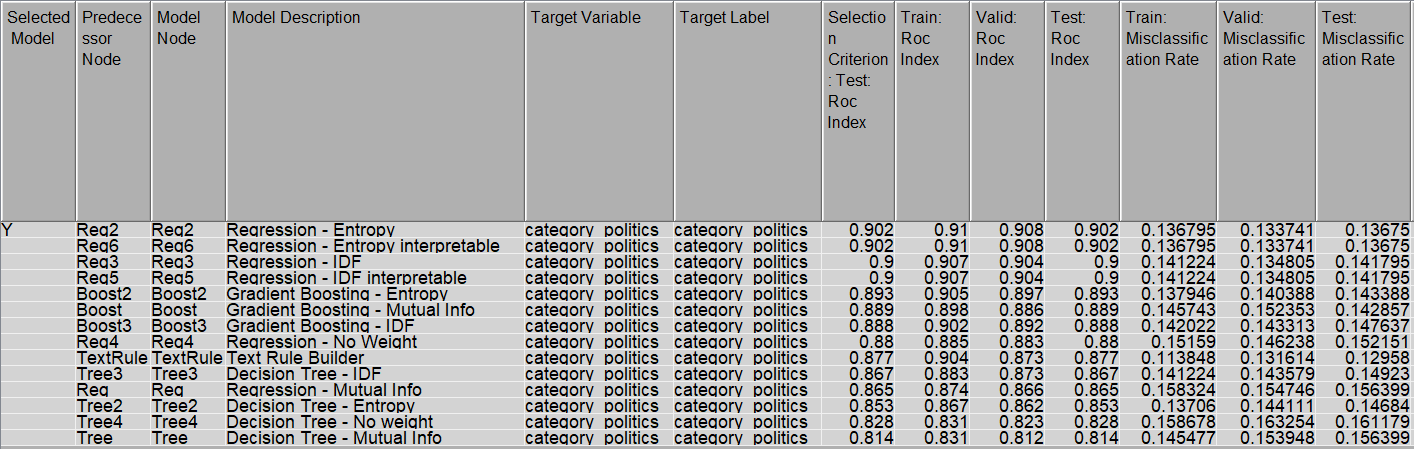
**Table 3: The Third Iteration Results**

In our fourth model iteration, we maintained the changes introduced in the third iteration but further streamlined the classification task by reducing the number of categories from 12 to 5. The motivation behind this adjustment was the hypothesis that a high classification rate might be influenced by the presence of too many classes. The selected categories for this model—culture, politics, business, health, and economy—were chosen based on their generality and being the most populated in the dataset. The reduction in the number of categories yielded significant improvements in the model's performance. While the ROC remained consistent at a high level of 0.911, there was a notable decrease in the Misclassification Rate, dropping to 0.318. This outcome suggests that simplifying the classification task by focusing on a smaller set of categories contributed positively to the overall effectiveness of our model.



**Table 4: The Fourth Iteration Results**

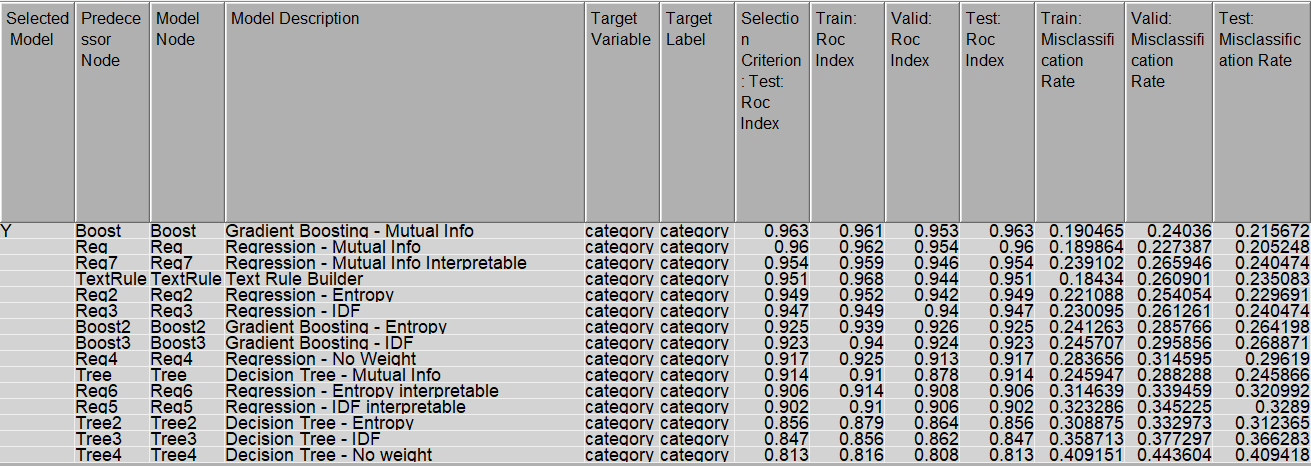
For the fifth iteration of our model, we experimented with generating dummy variables to transform the multi-class classification problem into a binary classification problem, aiming to address our misclassification rate. Up until the fourth iteration, the ROC remained relatively stable; however, there was no significant improvement in the misclassification rate. Consequently, we explored the conversion of our data to binary format to assess its impact on misclassification. To evaluate the effectiveness of this approach, we initiated the analysis with the 'politics' category. The model yielded a ROC of 0.902 and a misclassification rate of 0.137. Remarkably, this method successfully reduced the misclassification rate by more than 10%. Despite this improvement, it led to a slight drop in the ROC. Consequently, we made the decision to revert to the fourth iteration model but opted for a different category for further refinement.



**Table 5: The Fifth Iteration Results**

In our fourth iteration, even with a reduced set of five categories, we observed that health, economy, business, and politics remained interlinked or correlated in terms of the topics covered. This correlation among categories likely contributed to the previously high misclassification rate. To address this, we decided to explore the impact of selecting a different set of five categories in the sixth iteration.

In the sixth iteration, we strategically opted for a more diverse set of categories, including 'politics,' 'economy,' 'science,' 'music,' and 'sports,' each distinct from the others. The intent was to uncover insights that could enhance the development of a more comprehensive classification model. The outcomes of this iteration exceeded those of all previous attempts, achieving an impressive ROC of 0.961 and a reduced misclassification rate of 0.206. These results signify a significant advancement in the model's capability to accurately distinguish and categorize texts across a diverse range of topics. The strategic selection of categories in this iteration represents a crucial milestone in our iterative refinement process, underscoring the effectiveness of thoughtful category curation in improving overall model performance.

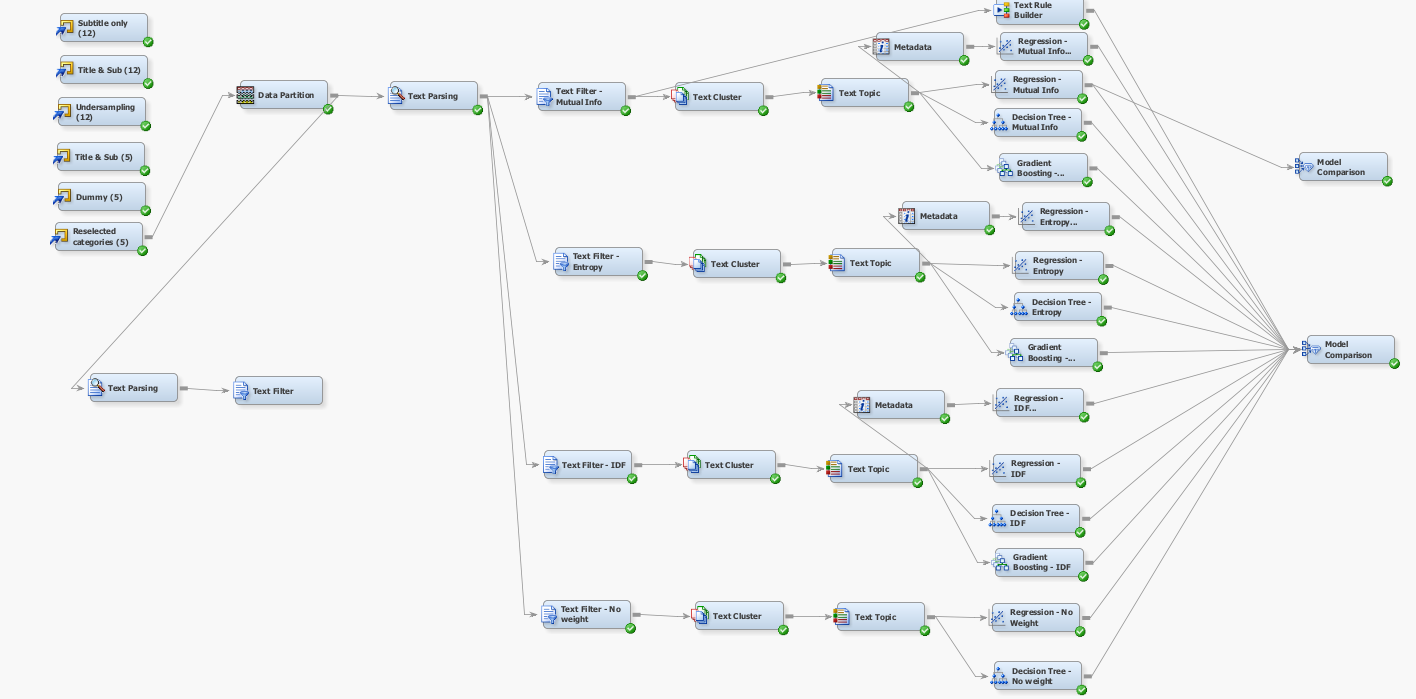


**Table 6: The Sixth Iteration Results**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Iter.  No. | Model | ROC | | | Misclassification Rate | | |
| Train | Validation | Test | Train | Validation | Test |
| 6th | Gradient Boosting - Mutual Info | 0.961 | 0.956 | 0.962 | 0.192 | 0.237 | 0.215 |
| 6th | Regression - Mutual Info | 0.963 | 0.958 | 0.961 | 0.197 | 0.226 | 0.206 |
| 4th | Regression - Mutual Info | 0.923 | 0.897 | 0.915 | 0.419 | 0.447 | 0.428 |
| 3rd | Regression - Mutual Info | 0.921 | 0.907 | 0.911 | 0.306 | 0.327 | 0.318 |
| 5th | Regression - Entropy | 0.918 | 0.895 | 0.911 | 0.412 | 0.429 | 0.429 |
| 2nd | Regression - Mutual Info | 0.910 | 0.908 | 0.902 | 0.137 | 0.134 | 0.137 |
| 1st | Regression - Mutual Info | 0.86 | 0.832 | 0.852 | 0.549 | 0.565 | 0.556 |

**Table 7: Comparison: The Best Results of Each Iteration**

**Final Model Interpretation:**



Throughout the entirety of our iterative refinement process, we employed a combination of regression, decision tree, and gradient boosting models. This comprehensive approach included variations in SVD resolution levels, dimensions, term weights, and frequency weighting. To facilitate a thorough comparison, interpretable models were incorporated for each type of term weighting method (for regression model only). As outlined in the preceding section, the culmination of our efforts resulted in the selection of the final model from the sixth iteration. This iteration, characterized by the inclusion of five distinct categories—'politics,' 'economy,' 'science,' 'music,' and 'sports'—yielded the most promising outcomes. The specifics of our final model are elaborated upon in the following section.

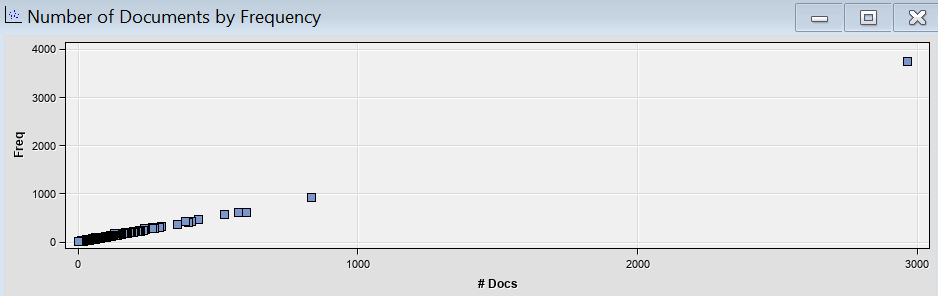
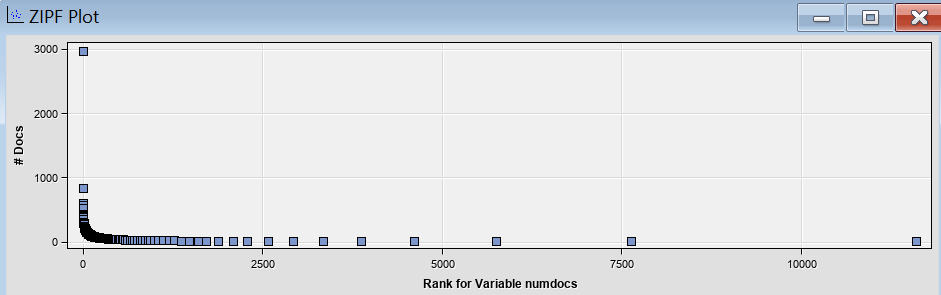
* **Summary of the Final Model:**

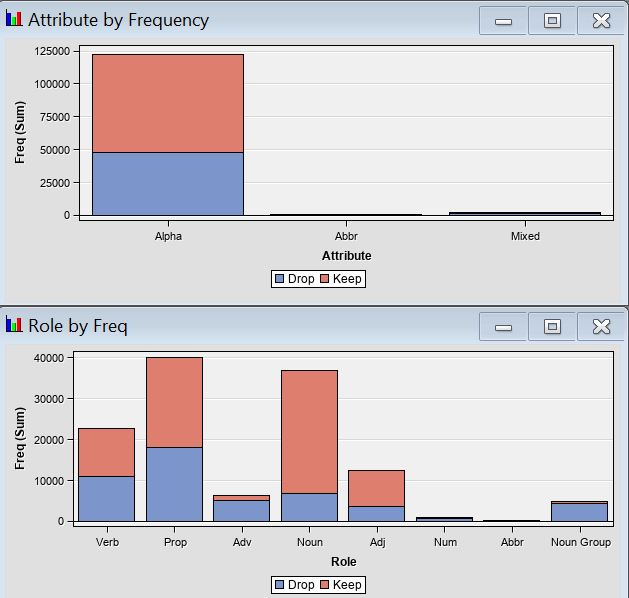
In our quest for an optimized text analysis framework, our final model boasts several key enhancements. We strategically divided our data into three segments with a ratio of 60:20:20, yielding a notable reduction in misclassification by ~5% compared to our initial attempt (40:30:30). Employing "log" for frequency weight and "Mutual Information" for term weight in the text filter, we set a low Singular Value Decomposition (SVD) resolution, capped at 100 dimensions, resulting in 46 SVDs. The culmination of 39 topics, comprising 25 auto-generated topics and 14 user-generated topics for the text topic, solidifies the versatility of our model.

|  |  |  |  |
| --- | --- | --- | --- |
| Data Partition | Train (60)  Valid (20)  Test (20) | Text Cluster | Resolution: low  max SVD dimension: 100  No. of SVD generated: 46 |
| Text Filter | Freq Weighting: Log  Term Weight: Mutual Information | Text Topic | Topic generated:  Auto generation: 25  User generation: 15 |

* **Text Filter Node Result:**

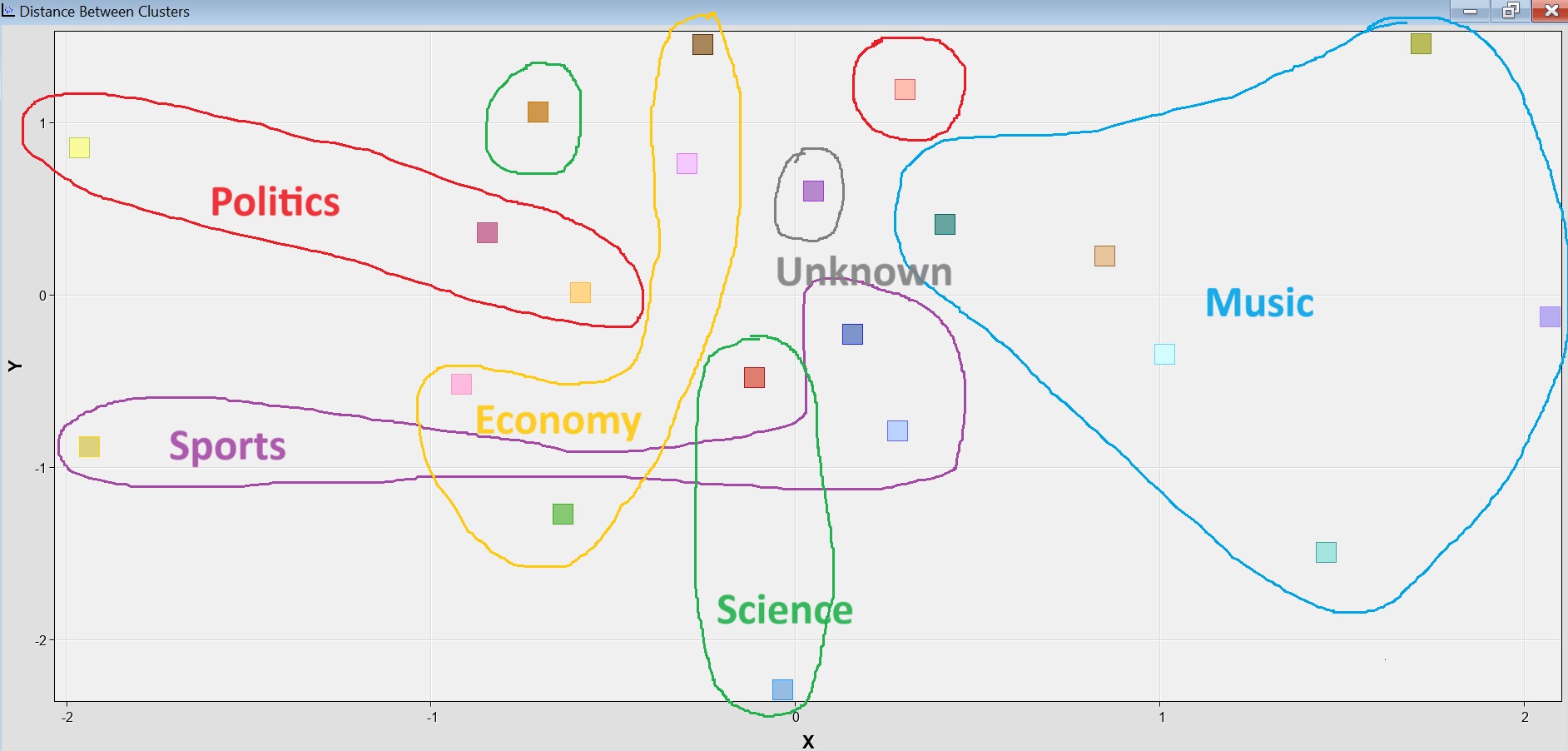
Analysis of the Zipf plot reveals exponential decay, and the linearity in the # of Docs by Frequency signifies a monotonic trend. Dominance by alpha terms, noun, verb, adjective and adverb suggests a robust data quality, assuring that our dataset doesn't encounter significant issues.





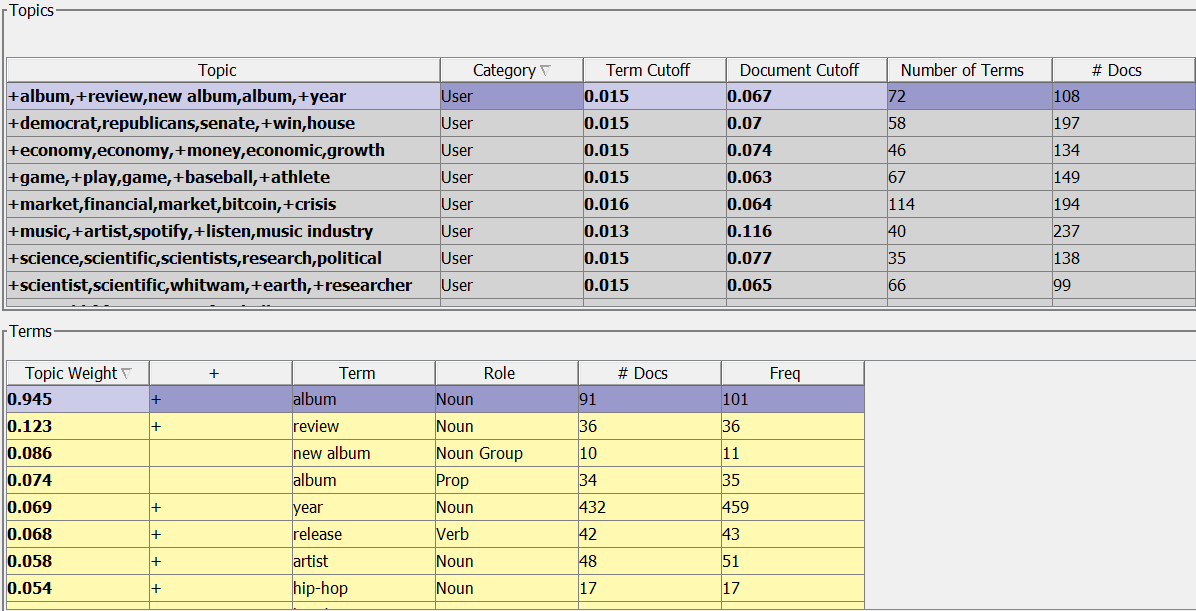
* **Text Cluster:**

The plotted distances between clusters unveil the Music category's distinctiveness, while the interrelatedness of the Economy, Politics, and Science categories is notable. Misclassification rates are somewhat explained by the challenge our model faces in accurately categorizing text containing terms situated in this specific area.



* **Text Topic:**

The table below showcases the topics generated alongside term weights. Experimentation with user-generated topics, achieved through term cutoff adjustments and topic weight modifications, contributed to a nuanced improvement in our model, though not a groundbreaking one.

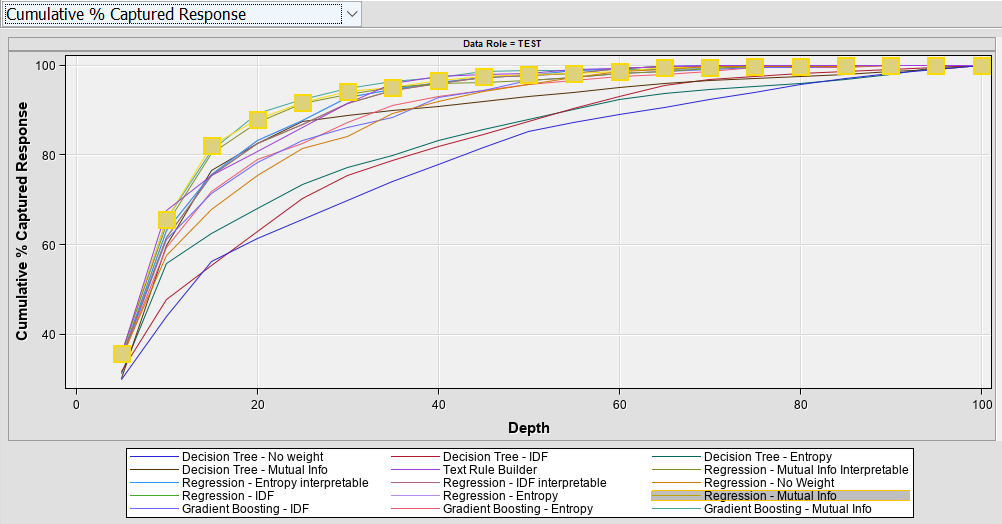


In examining the text topics, a noteworthy observation was made: a substantial number of non-politics-related topics prominently featured the term 'politic' with a significant topic weight. Recognizing the potential impact on our model's classification performance, a strategic decision was made to reduce the topic weight of 'politic' below the term cutoff values. This adjustment aimed to mitigate the undue influence of this term on non-politics topics. Moreover, when confronted with multiple terms deemed important but falling just short of the cutoff values by a marginal difference, we opted to adjust the term cutoff value itself rather than individually tweaking the topic weight of each term. This nuanced approach was implemented to maintain coherence and precision in our text topic analysis, contributing to the overall refinement of our model.

* **Cumulative % Captured Response:**

The cumulative captured response graph serves as a visual representation of our model's proficiency in identifying positive instances across various depths. Notably, our model excelled, achieving an 87.824% capture rate at depth 20 and attaining a nearly perfect 100% at depth 80. This underscores the effectiveness of our model in the early identification of a substantial proportion of relevant texts, indicative of a high recall rate.

While our model excels in identifying relevant texts early on, the graph also shows that it sometimes makes mistakes by marking incorrect instances as positive. Striking a balance is crucial — we want to catch as many relevant texts as possible (high recall rate) but also avoid including too many incorrect ones (minimizing false positives). Finding this balance is important for making our model more precise, ensuring that when it identifies positive instances, they genuinely match our intended criteria. Continuous fine-tuning and careful evaluation are key to refining our model's performance and maintaining a good balance between catching relevant texts and avoiding mistakes.



**Conclusion:**

The final model developed is able to relatively precisely categorize the posts on Medium.com with a misclassification rate of only 20%. There are, however, some limitations to this model. One such limitation comes from the information bias present in the data that was used. Medium.com is an online publication website that allows users to post stories about different topics. This skews the articles to the extremes on any topic since it is rare for a person with a moderate viewpoint to post something online. This skew changes the words that are used in the article titles and could result in a model that only works for extreme articles. Another limitation of the model is that in order to increase the ROC and decrease the misclassification rate, the number of target categories was reduced from 12 to only 5. This improved the performance of the model but decreased its capabilities of categorizing on a wider spectrum.

**Recommendations:**

After completing this text mining project, our team has developed some recommendations that would improve the flow of the project and final result were we to do it again. One recommendation would be to continue changing different parameters used to develop the model. For example, adding more terms to the Stop List, changing the term cutoff, and adding user topics could enhance the final model developed. The problem with these modification techniques is that they can take a long time to implement and can have an impact on each other which makes it virtually impossible to find the truly optimal solution.

Another suggestion would be to alter the problem statement slightly. Instead of trying to create a single model that will be able to categorize each post across a wide variety of categories, creating multiple narrower models might have been an interesting approach. This was something that we tried in our current project and is why we reached a model that can accurately categorize posts across 5 different categories ('politics', 'economy', 'science', 'music', 'sports') but that leaves 7 categories unaccounted for. If time had permitted, creating additional models might have yielded better results and would have allowed us to hit our target of being able to categorize any post made in a category with 1000+ entries. For example, making three models that cover four categories each would hit the target of 12 categories.

Finally, it might have been beneficial to do a deeper dive into the data that we were working with. This might have allowed us to find patterns in the data that we could exploit to build a better and more accurate model. We worked to understand the data and limit what was used so as to not pigeon hole ourselves by removing categories that were too specific such as ‘2020 Election’ and ‘Coronavirus’ but could have done more to develop links between certain phrases and words and their respective categories.

**Citations**

AMRRS. (2019). Medium Post Titles [Dataset]. Kaggle. https://www.kaggle.com/datasets/nulldata/medium-post-titles