Social Media Fake News Detection

Text Mining

OPIM 5671: Data Mining & Business Intelligence

Professor Sudip Bhattacharjee

TABLE OF CONTENTS

Introduction:

In today's digital age, where digital information reigns supreme, the internet and social media platforms serve as the main channels for global news consumption. However, this widespread accessibility has raised a critical issue – the credibility of online news. The surge in misinformation and the widespread prevalence of fake news presents significant risks, carrying extensive social and political implications. To address this growing challenge, this project aims to leverage advanced text mining techniques. The objective is to build a strong system capable of identifying and categorizing fake news articles, thereby playing a role in safeguarding the integrity of information in the digital era.

Problem Statement:

The internet, particularly driven by the widespread impact of social media, has ushered in an era where the prevalence of fake news has reached unprecedented levels. The unchecked dissemination of false information not only poses a threat to public perception but also carries significant consequences for decision-making processes and the fundamental structure of democratic societies. Given the immense volume of information online, manual verification becomes impractical and time-consuming. Hence, there arises a crucial requirement for an automated system capable of determining the credibility of news sources. Such a system is essential for providing users with reliable information and reinforcing the pillars of a credible and well-informed public discourse.

Software Used:

We employed SAS Enterprise Miner Workstation 15.1 to undertake our text mining initiative, benefiting from its advanced capabilities and sophisticated tools. The formalized environment provided by SAS Enterprise Miner facilitated a meticulous analysis of unstructured data, enabling us to extract meaningful insights with precision and methodological rigor.

Data Description:

Source: The data is sourced from news_articles.csv which we obtained from Kaggle.
Dimensions: The dataset comprises 2096 records and 12 fields.
Fields:

- **Author**: Refers to the author of the article.
- **Published:** Timestamp indicating the article's publication time.
- **Title**: Headline of the article.
- **Text**: The main body of the article.
- **Language**: Denotes the language in which the article is written.
- **Site_url**: Represents the website address where the article was published.

- ○ **Main_img_url**: URL of the primary image in the article.
- ○ **Type**: Indicates the category of the article (e.g., bias, fake, etc.).
- ○ **Label**: Designation of the article as either "Real" or "Fake."
- ○ **Title_without_stopwords:** The title text devoid of common words.
- ○ **Text_without_stopwords:** The main text without common words.
- ○ **Hasimage:** An indicator specifying if the article includes images.

- In this study, we conducted a comprehensive analysis of a dataset comprising 2097 records obtained from Kaggle, focusing on articles' authenticity.
- The dataset includes key attributes such as Author, Published timestamp, Title, Text, Language, Site_url, Main_img_url, Type (e.g., bias, fake), and Label (categorized as "Real" or "Fake"). To enhance the analysis, we processed the Title and Text by removing common stop words, resulting in Title_without_stopwords and Text_without_stopwords fields.
- Notably, the dataset exhibits a diverse range of news articles, with 672 records classified as Real news and 1190 records as Fake news, providing a substantial foundation for our investigation.
- For robust model development, we strategically split the data, allocating 50% for training, 30% for validation, and 20% for testing. This meticulous approach ensures a well-rounded evaluation of the model's performance. The entire dataset was meticulously cleaned, yielding 1863 records, which constitute the basis for our subsequent analysis and model development. The integration of these components lays the groundwork for a rigorous examination of factors contributing to the authenticity of news articles.

Snapshot of the Raw Data:

	A	B	C	D	E	F	G	H	I	J	K	L
1	author	published	title	text	language	site_url	main_img_	type	label	title_withc	text_withc	hasImage
2	Barracuda	2016-10-2	muslims bu	print they	english	100percen	http://bb4	bias	Real	muslims bu	print pay b	1
3	reasoning	2016-10-2	re why did	why did	english	100percen	http://bb4	bias	Real	attorney g	attorney g	1
4	Barracuda	2016-10-3	breaking w	red state	english	100percen	http://bb4	bias	Real	breaking w	red state f	1
5	Fed Up	2016-11-0	pin drop sp	email kayla	english	100percen	http://100	bias	Real	pin drop sp	email kayla	1
6	Fed Up	2016-11-0	fantastic t	email	english	100percen	http://100	bias	Real	fantastic t	email heal	1
7	Barracuda	2016-11-0	hillary goe	print	english	100percen	http://bb4	bias	Real	hillary goe	print hillar	1
8	Fed Up	2016-11-0	breaking n	breaking	english	100percen	http://100	bias	Real	breaking n	breaking n	1
9	Fed Up	2016-11-0	wow whist	breaking	english	100percen	http://100	bias	Real	wow whist	breaking n	1
10	Fed Up	2016-11-0	breaking cl	limbaugh	english	100percen	http://100	bias	Real	breaking cl	limbaugh s	1
11	Fed Up	2016-11-0	evil hillary	email	english	100percen	http://100	bias	Real	evil hillary	email peop	1
12	EdJenner	2016-11-0	yikes hillar	who	english	100percen	http://con	bias	Real	yikes hillar	comedian	1
13	Fed Up	2016-11-0	say goodby	students	english	100percen	http://100	bias	Real	say goodby	students e	1
14	EdJenner	2016-11-1	not kidding	email for	english	100percen	http://con	bias	Real	kidding col	email repu	1
15	Fed Up	2016-11-1	boom mat	copyright	english	100percen	http://100	bias	Real	boom mat	copyright p	1
16	Fed Up	2016-11-1	boom this	go to artic	english	100percen	http://100	bias	Real	boom pres	go article t	1
17	EdJenner	2016-11-1	trump supp	copyright	english	100percen	http://con	bias	Real	trump supp	copyright p	1

Data Preprocessing Steps and Justifications:

Field Reduction:

- **Removed Fields**: Published, Language, Site_url, Main_img_url, Type, Title_without_stopwords, Text_without_stopwords, Hasimage.

- **Justification**: Removal of these fields aims to focus the model on content-driven features. Fields like 'Published', 'Language', and 'Site_url' were deemed irrelevant to the authenticity of the content. Similarly, 'Main_img_url' and 'Hasimage' were considered less critical for text analysis. The 'Type' field was redundant given the presence of the 'Label' field.

Combining Title and Text:

- **Action Taken:** The 'Title' and 'Text' fields were amalgamated into a single ' Combined_Text' field**.**

- **Justification:** Combining these fields allows the model to consider the interplay between the title and the body, enhancing its ability to discern nuances in the news content.

Cleaning Data:

- **Action Taken:** The dataset underwent cleaning to eliminate records with missing or incomplete information, resulting in a reduction from 2096 to 1863 records.

- **Justification:** Cleaning ensures the model trains on quality, complete data, thereby improving its predictive accuracy and reliability.

In summary, the preprocessing steps undertaken for the social media Fake News Detection model were crucial in optimizing the data for training a reliable and effective model. This documentation provides transparency and clarity on the rationale and methods employed in the preprocessing phase, ensuring a thorough understanding of the data preparation process.

Snap shot of the Processed Data:

author	label	combined_text
Barracuda Brigade	Real	muslims busted they stole millions in govt benefits print they should pay all the back all the money plus interest the entire family and everyone who came in with them need to be deported asap wh
reasoning with facts	Real	re why did attorney general loretta lynch plead the fifth why did attorney general loretta lynch plead the fifth barracuda brigade print the administration is blocking congressional probe into cash pa
Barracuda Brigade	Real	breaking weiner cooperating with fbi on hillary email investigation red state fox news sunday reported this morning that anthony weiner is cooperating with the fbi which has reopened yes lefties r
Fed Up	Real	pin drop speech by father of daughter kidnapped and killed by isis i have voted for donald j trump percentfedupcom email kayla mueller was a prisoner and tortured by isis while no chance of relea
Fed Up	Real	fantastic trumps point plan to reform healthcare begins with a bombshell percentfedupcom email healthcare reform to make america great again since march of the american people have had to
Barracuda Brigade	Real	hillary goes absolutely berserk on protester at rally video print hillary goes absolutely berserk she explodes on bill rapist protester at rally oh the irony she is an enabler to bills escapades shes is just
Fed Up	Real	breaking nypd ready to make arrests in weiner casehillary visited pedophile island at least timesmoney laundering underage sex payforplayproof of inappropriate handling classified information pe
Fed Up	Real	wow whistleblower tells chilling story of massive voter fraud trump campaign readies lawsuit against fl sec of elections in critical district video percentfedupcom breaking nypd ready to make arres
Fed Up	Real	breaking clinton clearedwas this a coordinated last minute trick to energize hillarys base percentfedupcom limbaugh said that the revelations in the wikileaks material were starting to hurt the clin
Fed Up	Real	evil hillary supporters yell fck trumpburn truck of daddy fishing with yr son over of trump bumperstickers video percentfedupcom email these people are sick and evil they will stop at nothing to ge
EdJenner	Real	yikes hillary goes off the railspulls a howard dean video who comedian where would she move spain i did buy a house in another country just in case so all of these people that threaten to leave th
Fed Up	Real	say goodbye these hollywood celebs threatened to leave the uslets hold them to it percentfedupcom students expressed their fear over a trump presidency in messages to each other that were b
EdJenner	Real	not kidding colleges give students safe spaces to cry over trump winthreaten students over protrump chalkings email for republican politicians like ohio governor john kasich who refused to get beh
Fed Up	Real	boom math shows trump would have beaten obama in romneyobama election percentfedupcom copyright percentfedupcom in association with liberty alliance all rights reserved proudly built b
Fed Up	Real	boom this is how president reagan handled protesters negotiate what is there to negotiate video percentfedupcom go to article a trump supporter wearing a trumppence tshirt let it fly on a repor
EdJenner	Real	trump supporter got nuts on msnbc reporter covering antitrump rioters video copyright percentfedupcom in association with liberty alliance all rights reserved proudly built by wpdevelopers stay
Fed Up	Real	tomi lahren has special message for celebrities who said theyd move to canada if trump won video percentfedupcom go to article donald trump was willing to give up a very fulfilling life that took
EdJenner	Real	boycottcomedianrobert deniro wanted to punch trump in the facesupports antitrump riotersnow wants americans to support his new movie video john mcnaughton is a special american painter b
EdJenner	Real	hes never sold an original painting until nowand this ones going in the white house go to article dear abby i supported a woman i knew had a history of criminal activity who is married to a rapist ar
EdJenner	Real	sorry liberalsyou can stop with the petitionshillary did not win the popular vote mark cuban has made no secret of his dislike for trump and his love for crooked hillary watch him tell fox news neil c
Fed Up	Real	mark cuban in the event donald wins i have no doubt the market tanksso heres what really happened video percentfedupcom david wilcox a year old chicago man who was brutally beaten by a m

2. Text Mining - Important Components:

- **Tokenization:** Tokenization is the process of breaking text into individual units called tokens. Tokens can be words, sentences, or even smaller units like characters or n-grams. Tokenization is a fundamental step in text mining and natural language processing (NLP) tasks.

- **Stop Words:** Stop words are common words that are often removed from text during preprocessing because they do not carry significant meaning. Examples of stop words include "the," "is," "and," and "in." Removing stop words can help reduce noise and improve the efficiency of text mining algorithms.

- **Stemming and Lemmatization:** Stemming and lemmatization are techniques used to reduce words to their base or root forms. Stemming involves removing affixes from words, resulting in a truncated version. Lemmatization, on the other hand, transforms words to their canonical or dictionary forms. Both techniques help normalize and group similar words together for analysis.

- **Term Frequency-Inverse Document Frequency (TF-IDF):** TF-IDF is a numerical representation of the importance of a term within a document or a corpus. It considers both the frequency of a term in a document (TF) and its rarity across the entire corpus (IDF). TF-IDF is commonly used for text classification, information retrieval, and keyword extraction.

- **Term-Document Matrix:** A term-document matrix (TDM) is a representation of a corpus of documents in which each row corresponds to a unique term (word) in the corpus, and each column corresponds to a document. The entries of the matrix represent the frequency or presence of the term in each document.

- **Sentiment Analysis:** Sentiment analysis aims to determine the emotional tone or sentiment expressed in a piece of text. It can involve classifying text as positive, negative, or neutral, or assigning sentiment scores to indicate the intensity of positive or negative sentiment. Sentiment analysis is used in various applications like customer feedback analysis and social media monitoring.

2.1 Text Mining: Modeling and Forecasting - Important Components

- **Test ROC Index:** Test ROC Index is a performance measure used in binary classification tasks. It quantifies the ability of a classification model to discriminate between positive and negative instances by plotting the true positive rate against the false positive rate.

- **Test Misclassification Rate:** In text mining, test misclassification refers to the rate at which instances are incorrectly classified by a classification model. It measures the proportion of misclassified instances, indicating the accuracy of the model in predicting the correct class labels for text data.

- **Prediction Errors:** In text mining, prediction errors refer to the discrepancies between the actual values or labels of the text instances and the predicted values or labels assigned by a text mining model. These errors quantify the differences between the model's predictions and the ground truth values.

- **Mean Absolute Percentage Error:** MAPE can be used as an error metric to evaluate the performance of prediction models. It calculates the average absolute difference between the actual values or labels of the text instances and the predicted values or labels, divided by the actual values or labels, expressed as a percentage.

- **AIC:** AIC can be used in text mining to assess the goodness-of-fit of models without overfitting. It rewards models that achieve a high level of fit to the data while penalizing overly complex models.

- **SBC:** BIC, also known as Schwarz Information Criterion (SIC) or SBC, is another model selection criterion used in text mining. Similar to AIC, it considers the likelihood function and penalizes complex models. Lower BIC values indicate preferred models.

- **RMSE:** RMSE can be employed as a metric to evaluate the accuracy of text mining models. It calculates the square root of the mean of the squared differences between the actual values or labels and the predicted values or labels. RMSE provides an indication of the average magnitude of the prediction errors.

3. Nodes Used in SAS Studio:

- **File Import:** SAS Enterprise Miner Workstation allows you to import text data from various file formats, such as plain text files, Microsoft Word documents, PDFs, or web pages. You can use the built-in data import capabilities to bring your text data into the tool for further analysis.

- **Data Partition:** SAS Enterprise Miner Workstation provides options to partition your text data into training and testing sets. You can easily split your dataset into subsets for model development and evaluation purposes. This helps in ensuring that your models are trained on a portion of the data and tested on an unseen portion for unbiased performance evaluation.

- **Text Parsing:** SAS Enterprise Miner Workstation offers text parsing functionalities to preprocess and parse text data. You can tokenize text into words or other linguistic units, segment sentences, perform part-of-speech tagging, and conduct syntactic parsing. These features assist in extracting structured information from unstructured text data.

- **Text Filtering:** SAS Enterprise Miner Workstation provides options for text filtering and preprocessing. You can apply various filters to remove stop words, punctuation, special characters, or other unwanted elements from your text data. These filtering techniques help clean and prepare your text data for further analysis.

- **Text Clustering:** SAS Enterprise Miner Workstation includes clustering algorithms for text data. You can apply these algorithms to group similar documents together based on their content. The clustering capabilities help you identify patterns, themes, or topics in your text data by organizing related documents into meaningful clusters.

- **Text Topic:** SAS Enterprise Miner Workstation supports topic modeling techniques for extracting topics from text data. You can utilize algorithms like Latent Dirichlet Allocation

(LDA) to identify the underlying topics present in your text corpus. The topic modeling capabilities enable you to gain insights into the main subjects discussed in your text data.

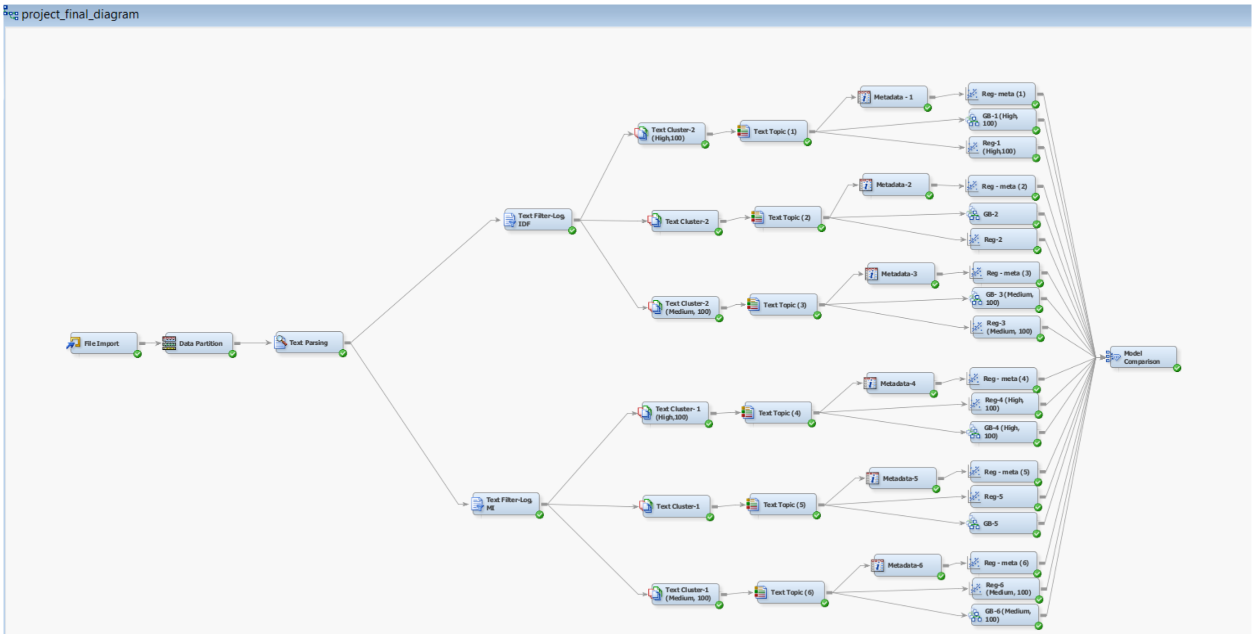
- **Model Comparison:** SAS Enterprise Miner Workstation offers tools to compare and evaluate different text mining models. You can assess and compare the performance of various models using evaluation metrics like Test Roc Score, Misclassification rate, RMSE, accuracy, precision, recall, F1-score, or other domain-specific measures. Model comparison helps in selecting the most suitable model for your text mining task.

- **Metadata:** SAS Enterprise Miner Workstation allows you to work with metadata associated with your text data. You can incorporate metadata attributes such as author, publication date, source, document type, or any other relevant information into your analysis. Metadata provides additional context and details about your text corpus, enhancing the understanding and interpretation of the data.

- **Scoring:** SAS Enterprise Miner Workstation enables you to assign scores to documents or text instances based on certain criteria or models. You can score documents for ranking, sentiment analysis, relevance assessment, or other purposes. The scoring capabilities provide a quantitative measure to aid decision-making and further analysis.

FULL MODEL DIAGRAM

Please find the below final model for our project:



MODEL DESCRIPTION:

File Import:

We have imported the data and set the roles for the variables.

Name	Role	Level
author	**Rejected**	**Nominal**
combined_text	**Text**	**Nominal**
label	**Target**	**Binary**

Since there is no significance for the author in detecting the output, we have rejected it. Label is our target variable and combined text is our text variable.

Data Partition:

We have split the data into 3 partitions for interpreting the model:
Training – 50
Validation – 30
Test – 20

Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	50.0
Validation	30.0
Test	20.0

Data Partition Node Results:

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
label	.		1	0.0537	label
label	.	Fake	1190	63.8755	label
label	.	Real	672	36.0709	label

Data=TEST

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
label	.	Fake	239	64.0751	label
label	.	Real	134	35.9249	label

Data=TRAIN

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
label	.		1	0.1074	label
label	.	Fake	594	63.8024	label
label	.	Real	336	36.0902	label

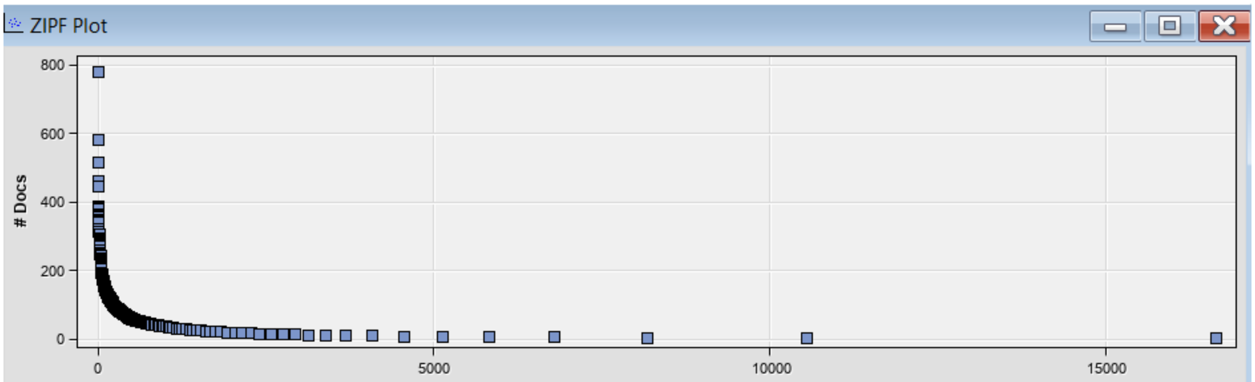
Data=VALIDATE

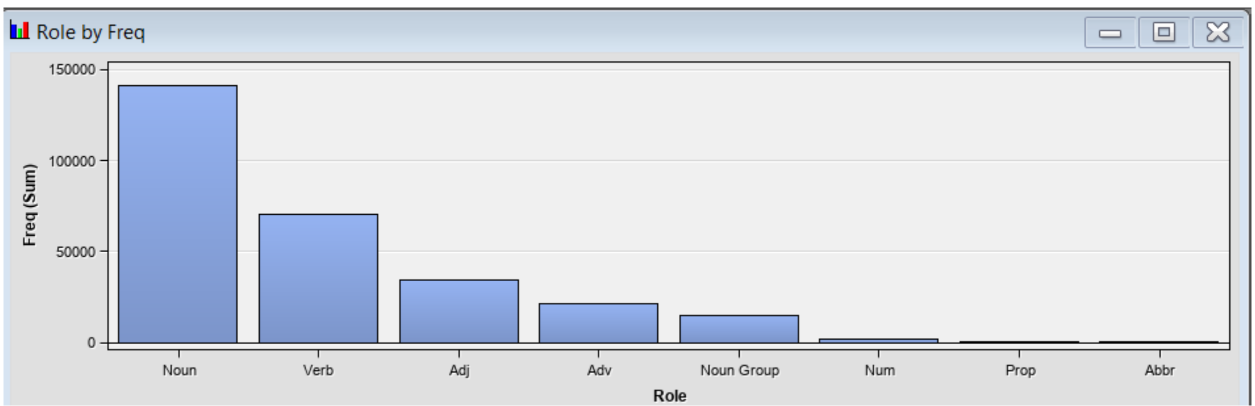
Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
label	.	Fake	357	63.8640	label
label	.	Real	202	36.1360	label

Text Parsing:

Property	Value
General	
Node ID	TextParsing
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Parse	
Parse Variable	combined_text
Language	English
Detect	
Different Parts of Speec	Yes
Noun Groups	Yes
Multi-word Terms	SASHELP.ENG_MULT
Find Entities	None
Custom Entities	
Ignore	
Ignore Parts of Speech	'Aux' 'Conj' 'Det' 'Inte
Ignore Types of Entities	...
Ignore Types of Attribut	'Num' 'Punct'
Synonyms	
Stem Terms	Yes
Synonyms	SASHELP.ENGSYNMS
Filter	
Start List	...
Stop List	SASHELP.ENGSTOP
Select Languages	...
Report	
Number of Terms to Di:	20000

In Text Parsing, we have used the default settings for all the attributes.

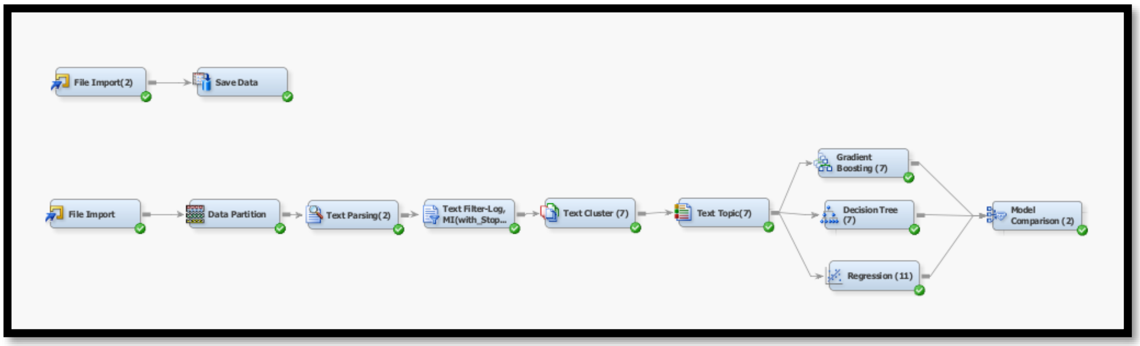




We have manually created a stop list that excludes words from the documents that appear rarely in just some of the documents or if they appear in almost all the documents since these two kinds doesn't have much of significance and tried used different supervised learning techniques on them.

But the results were not satisfactory. They are having misclassification rate higher than the models that are using the default stop list.

The results are as below:

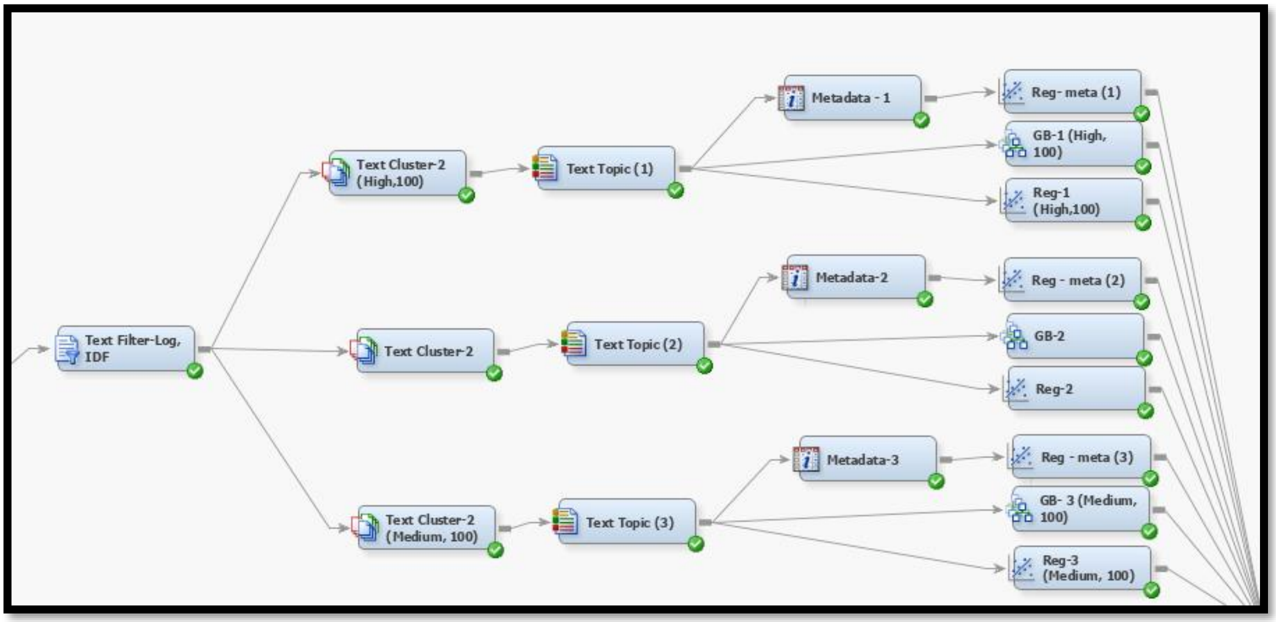


Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Test: Misclassification Rate	Test: Roc Index ▼
	Reg11	Reg11	Regression (11)	label	label	0.320215	0.284182	0.776
Y	Boost7	Boost7	Gradient Boosting (7)	label	label	0.26297	0.27882	0.766
	Tree7	Tree7	Decision Tree (7)	label	label	0.325581	0.33244	0.686

The results are not favorable to the model. Though they have a decent ROC, it is more compared to our best model. So, we have excluded this from our final model.

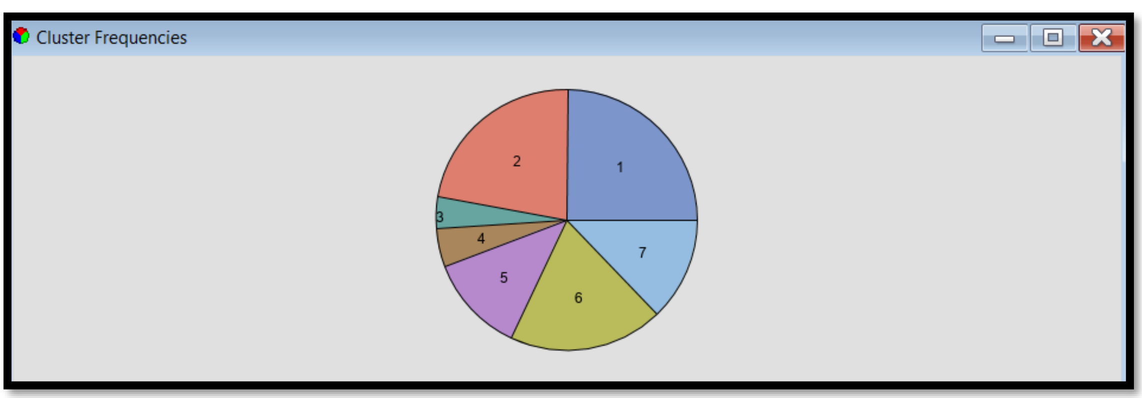
Text Filter with Frequency Weighting Log and Term Weight Inverse Document Frequency (IDF):



To determine the most effective frequency weight option for our analysis, we experimented with various options in the Text Filter node of SAS Enterprise Miner Workstation. First we tried with frequency weight as Log, Binary but observed that Log is giving better results compared to Binary. So, in the first combination we have taken Frequency weight as Log and Term weight as Inverse Document Frequency (IDF). We have left all the values to be default.

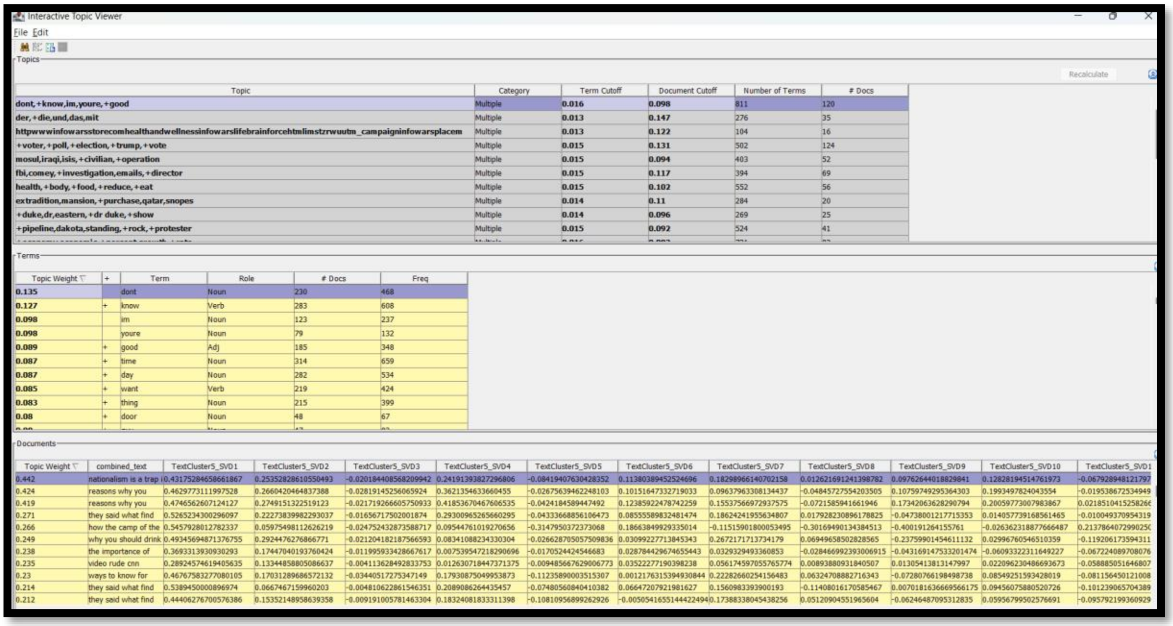
Weightings	
Frequency Weighting	Log
Term Weight	Inverse Document Frequency

Text Cluster with SVD Resolution High and Max SVD Dimensions 100:



Cluster ID	Descriptive Terms	Frequency	Percentage	Coordinate 1	Coordinate 2	Coordinate 3	Coordinate 4	Coordinate 5	Coordinate 6	Coordinate 7	Coordinate 8	Coordinate 9	Coordinate 10	Coordinate 11	Coordinate 12
1	political +power +world +mean +man first +include +place +back +article +find +year +work +point +me...	230	25%	0.399572	-0.00376	0.001833	-0.01486	-0.02312	-0.00179	.0003391	-.000222	0.008433	-0.0147	-.000584	0.02197
2	+trump donald +vote +election hillary +win +'donald trump' +poll republican +candidate +voter +supporter ...	210	23%	0.480971	-0.13175	-0.01597	0.044387	-0.14098	0.0304	0.043715	0.118481	0.098514	-0.04632	0.091417	0.145749
3	+die der auf mit und das sich ein nicht zu +hat auch den für ist ...	35	4%	0.034063	-0.0388	-0.99159	-0.0387	0.040573	-0.01698	-0.00536	-0.0052	0.005321	0.00218	-0.01381	-0.00857
4	+food water le naturalnews gorafi natural health +tag +eat +contain +source +avoid +body +help +cause...	46	5%	0.310042	0.134439	-0.00112	0.0207	0.152961	-0.11862	0.054401	0.116189	-0.17869	-0.01343	-0.01566	0.035729
5	clinton clintons +investigation fbi emails hillary +director 'hillary clinton' +break +campaign democratic em...	113	12%	0.42267	-0.29849	0.014088	0.026092	-0.05446	-0.08942	0.016398	0.075995	0.032253	0.064692	-0.03469	0.004425
6	+body +place +cause health dont +good +help natural +find +know +long +day +life +eat +big ...	179	19%	0.465309	0.112216	-0.00961	0.04915	0.02596	-0.04484	0.068319	0.033711	-0.05285	0.00372	-0.00172	0.012582
7	military syria +photo +terrorist loading +group foreign +city russia +force +government +month +official ...	118	13%	0.377659	-0.03534	0.00573	-0.01798	-0.00121	-0.01086	-0.15455	-0.11461	-0.11028	-0.11449	0.090166	-0.04659

For this Configuration, the data is divided into 7 clusters and 25 topics with different Document cutoff and Term cutoff.



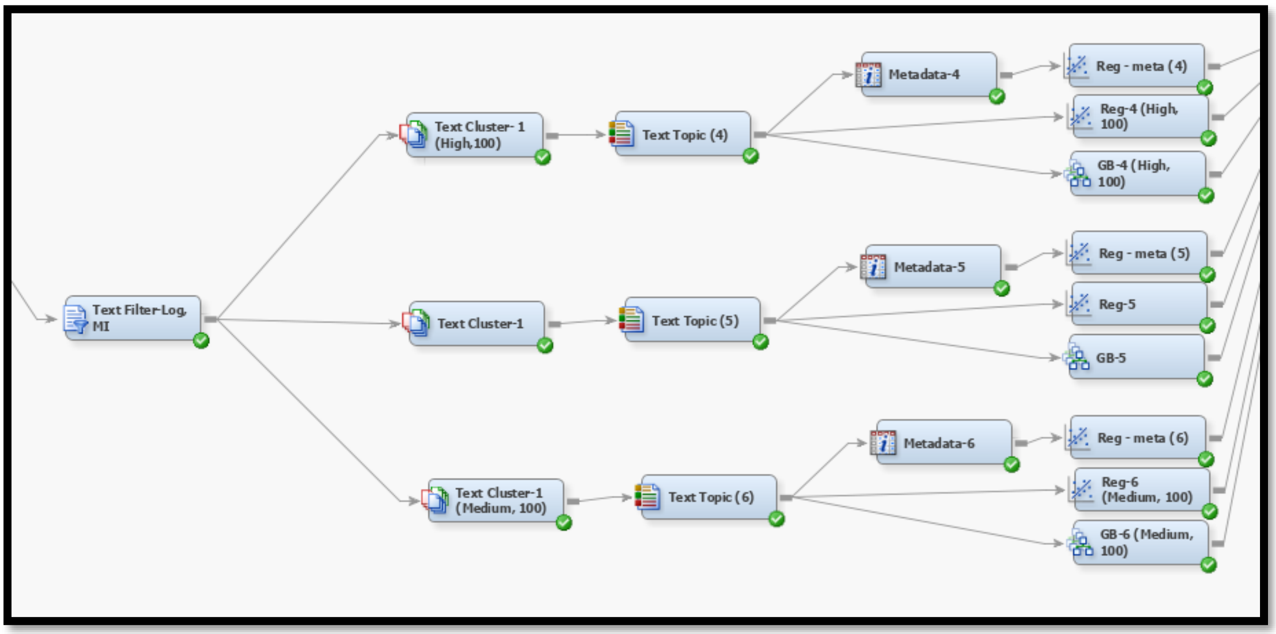
In a similar way, we have performed different SVD dimensions of SVD Resolution Low, Max SVD Dimensions 100 and SVD Resolution Medium, Max SVD Dimensions 100.

For these we have applied Supervised learning techniques such as Gradient Boosting and Regression. The following are the parameters we got for these models.

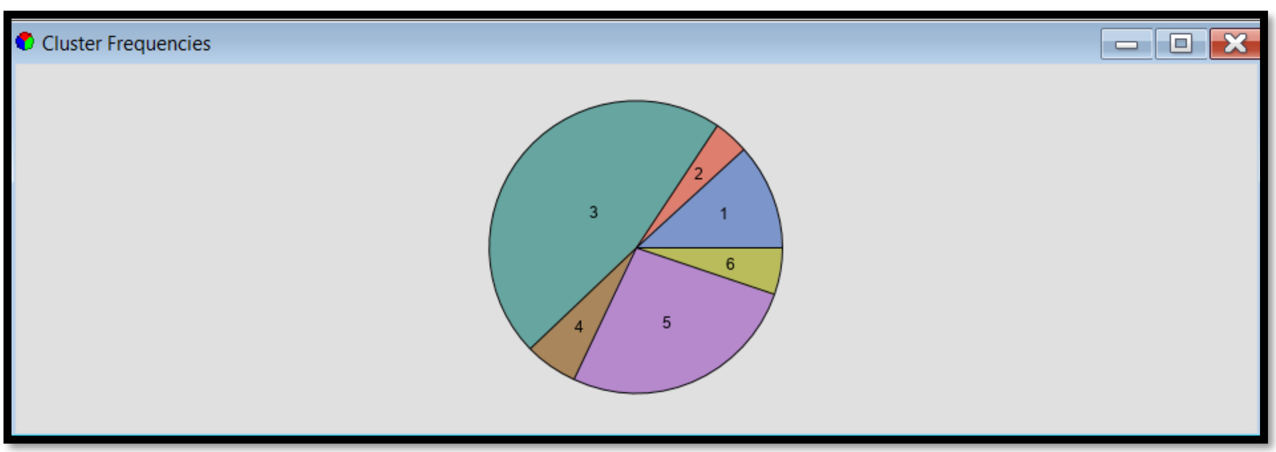
	SVD values	Model Type	Test: ROC Index
Term Weight IDF	High, 100	Regression	0.784
		GB	0.797
	Low, 100	Regression	0.788
		GB	0.777
	Medium, 100	Regression	0.794
		GB	0.775

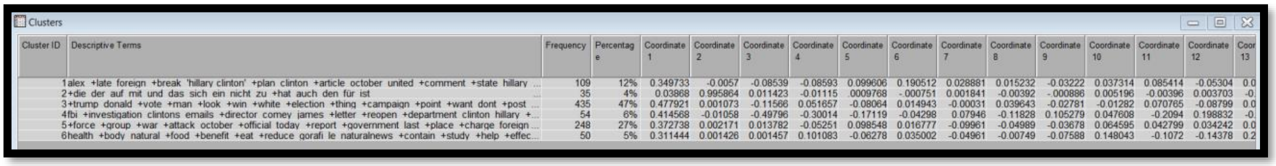
Text Filter with Frequency Weighting Log and Term Weight Mutual Information (MI):

Generally, for any data consisting of the target variable Mutual Information gives the best results, however performing IDF doesn't cause any harm which we have performed previously. Since our dataset has a target variable **Label,** we are opting for Mutual Information now.

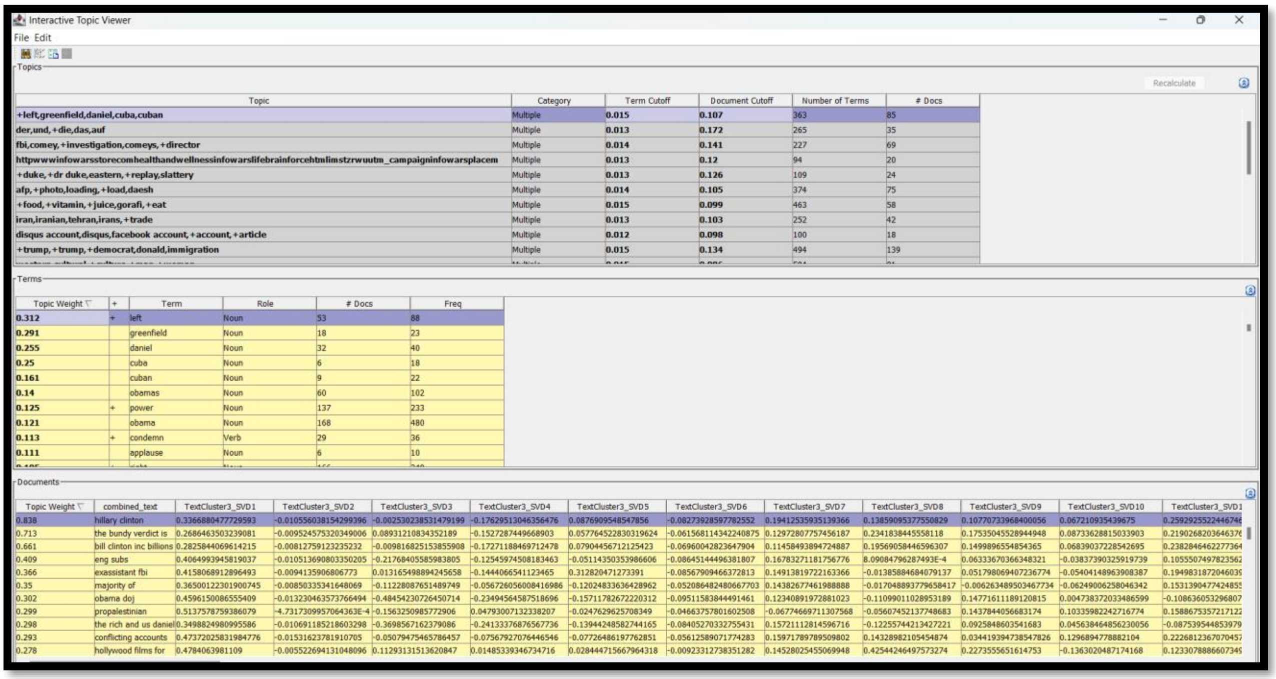


Text Cluster with SVD Resolution Medium and Max SVD Dimensions 100:





These results are for SVD Resolution Medium and Maximum Dimensions to be 100. We got 7 clusters and 25 Text Topics. We have examined the words and included. Excluded them by changing the Term Cutoff and Document cutoff.

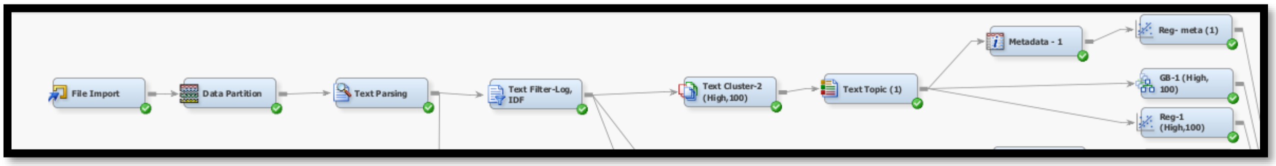


We have also experimented with SVD Resolution Low and High. In a similar way as IDF, we tried different supervised learning techniques such as Regression, Gradient Boosting for these combinations.

The below are the results:

	SVD values	Model Type	Test: ROC Index
Term Weight Mutual Information	High, 100	Regression	0.824
		GB	0.817
	Low, 100	Regression	0.821
		GB	0.809
	Medium, 100	Regression	0.829
		GB	0.819

INTERPRETABLE MODEL



We've opted not to move forward with all the variables due to possible interpretation challenges. Consequently, we've excluded SVD inputs and raw text topic from our analysis. Our focus now centers on utilizing text topic and text cluster probability.

Interpretable models enhance user trust by providing clear insights into predictions, ensuring compliance with transparency regulations. They simplify error diagnosis, reducing the likelihood of systematic errors, and offer valuable insights into feature importance. In model improvement, interpretable models guide enhancements, informing feature engineering and model selection. Additionally, they foster effective communication in interdisciplinary settings, promoting collaboration between machine learning experts and non-technical stakeholders.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	-6.9275	59.3930	0.01	0.9071		0.001
TextCluster5_cluster_	1	1	2.1173	11.4706	0.03	0.8536		8.308
TextCluster5_cluster_	2	1	-0.6155	11.5632	0.00	0.9575		0.540
TextCluster5_cluster_	3	1	-10.2112	68.4155	0.02	0.8814		0.000
TextCluster5_cluster_	4	1	3.5930	12.2728	0.09	0.7697		36.342
TextCluster5_cluster_	5	1	0.6923	11.7946	0.00	0.9532		1.998
TextCluster5_cluster_	6	1	2.4898	11.4879	0.05	0.8284		12.059
TextCluster5_prob1		1	-0.6590	2.9094	0.05	0.8208	-0.1516	0.517
TextCluster5_prob2		1	2.3634	3.4208	0.48	0.4896	0.5352	10.627
TextCluster5_prob3		0	0	.	.	.	.	.
TextCluster5_prob4		1	-3.8597	6.1482	0.39	0.5301	-0.4547	0.021
TextCluster5_prob5		1	0.7109	4.3871	0.03	0.8713	0.1270	2.036
TextCluster5_prob6		1	-1.7928	3.0091	0.35	0.5513	-0.3734	0.167
TextCluster5_prob7		0	0	.	.	.	.	.
TextTopic2_1	0	1	-0.0352	0.1304	0.07	0.7874		0.965
TextTopic2_10	0	1	-0.3674	0.2049	3.22	0.0729		0.693
TextTopic2_11	0	1	0.1353	0.1454	0.87	0.3522		1.145
TextTopic2_12	0	1	-0.2770	0.1502	3.40	0.0651		0.758
TextTopic2_13	0	1	0.1330	0.1596	0.69	0.4049		1.142
TextTopic2_14	0	1	-0.2057	0.1528	1.81	0.1783		0.814
TextTopic2_15	0	1	-0.2814	0.1206	5.44	0.0196		0.755
TextTopic2_16	0	1	0.0497	0.1605	0.10	0.7566		1.051
TextTopic2_17	0	1	-0.2660	0.1337	3.96	0.0466		0.766
TextTopic2_18	0	1	-0.1703	0.1410	1.46	0.2270		0.843
TextTopic2_19	0	1	0.3502	0.1695	4.27	0.0389		1.419
TextTopic2_2	0	0	0	.	.	.	.	.
TextTopic2_20	0	1	-0.0240	0.1760	0.02	0.8916		0.976
TextTopic2_21	0	1	-0.0514	0.1306	0.16	0.6936		0.950
TextTopic2_22	0	1	-0.3200	0.1416	5.11	0.0238		0.726
TextTopic2_23	0	1	0.2954	0.2027	2.13	0.1449		1.344
TextTopic2_24	0	1	-0.1182	0.1640	0.52	0.4711		0.889
TextTopic2_25	0	1	0.0275	0.1236	0.05	0.8238		1.028
TextTopic2_3	0	1	5.7953	58.2421	0.01	0.9207		328.761
TextTopic2_4	0	1	0.3877	0.1329	8.51	0.0035		1.474
TextTopic2_5	0	1	-0.0795	0.1907	0.17	0.6768		0.924
TextTopic2_6	0	1	-0.4881	0.1734	7.93	0.0049		0.614
TextTopic2_7	0	1	1.0332	0.3947	6.85	0.0088		2.810
TextTopic2_8	0	1	0.2826	0.3158	0.80	0.3708		1.327
TextTopic2_9	0	1	-0.9810	0.2883	11.58	0.0007		0.375

Model	**Model Description**	**Test: ROC Index**	**Test: Misclassification Rate**
Regression	Reg-meta (1)	0.721	0.329

8.Model Comparison:

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Test: Misclassification Rate	Test: Roc Index ▼
	Reg7	Reg7	Reg-6 (Medium, 100)	label	label	0.288014	0.235925	0.829
	Reg6	Reg6	Reg-4 (High, 100)	label	label	0.264758	0.235925	0.824
	Reg	Reg	Reg-5	label	label	0.257603	0.246649	0.82
Y	Boost3	Boost3	GB-6 (Medium, 100)	label	label	0.250447	0.24933	0.819
	Boost2	Boost2	GB-4 (High, 100)	label	label	0.264758	0.254692	0.817
	Boost	Boost	GB-5	label	label	0.261181	0.217158	0.809
	Boost5	Boost5	GB-1 (High, 100)	label	label	0.307692	0.273458	0.797
	Reg10	Reg10	Reg-3 (Medium, 100)	label	label	0.280859	0.273458	0.794
	Reg4	Reg4	Reg - meta (5)	label	label	0.284436	0.27882	0.793
	Reg8	Reg8	Reg-2	label	label	0.296959	0.270777	0.788
	Reg9	Reg9	Reg-1 (High,100)	label	label	0.284436	0.294906	0.784
	Reg5	Reg5	Reg - meta (6)	label	label	0.271914	0.254692	0.783
	Boost4	Boost4	GB-2	label	label	0.293381	0.281501	0.777
	Boost6	Boost6	GB- 3 (Medium, 100)	label	label	0.288014	0.284182	0.775
	Reg3	Reg3	Reg - meta (4)	label	label	0.277281	0.270777	0.756
	Reg12	Reg12	Reg- meta (1)	label	label	0.304114	0.329759	0.731
	Reg2	Reg2	Reg - meta (3)	label	label	0.302326	0.324397	0.691
	Reg13	Reg13	Reg - meta (2)	label	label	0.313059	0.324397	0.679

Suitability of AIC for Binary Output and Imbalanced Data:

- Binary Output Consideration:
  - Models dealing with binary output (news as real or fake) often face challenges in balancing model complexity with predictive accuracy. AIC effectively addresses this by penalizing unnecessary complexity while rewarding good fit to the data.
- Handling Imbalanced Data:
  - In scenarios with imbalanced datasets (disproportionate number of real vs. fake news articles), models might lean towards the majority class. AIC helps in

comparing models not just on their accuracy but based on how well they explain the data, which is crucial in imbalanced situations.

- Emphasis on True Positives and False Positives:
  - AIC indirectly considers true positives and false positives. A model that generates many false positives or misses many true positives will have a poorer fit to the data, resulting in a higher AIC.
  - This characteristic of AIC makes it a suitable criterion for models where the correct classification of both classes (real and fake news in this case) is equally important.

- On the basis of ROC, we identified Regression 6, having SVD resolution as Medium and SVD dimension as 100 having ROC value 0.829 and 76.5% accuracy to be the best model compared to others.

Final Best Model:

Regression 6 model

Outputs:

TextTopic7_raw4	1	-28.9880	24.6067	1.39	0.2388	-1.6178	0.000
TextTopic7_raw5	1	3.2885	3.2912	1.00	0.3177	0.1769	26.801
TextTopic7_raw6	1	18.4846	4.9864	13.74	0.0002	0.7028	999.000
TextTopic7_raw7	1	-12.0161	11.9802	1.01	0.3159	-0.4572	0.000

TextTopic7_raw12	1	1.4423	3.8319	0.14	0.7066	0.0546	4.231
TextTopic7_raw13	1	5.4569	3.5053	2.42	0.1195	0.1816	234.379
TextTopic7_raw14	1	-43.6230	13.1824	10.95	0.0009	-1.6035	0.000
TextTopic7_raw15	1	2.6280	3.6588	0.52	0.4726	0.0962	13.845
TextTopic7_raw16	1	1.9504	3.3050	0.35	0.5551	0.0685	7.032

Positive Predictors: The model identified a specific cluster, labeled as `TextTopic7_raw6`, which is strongly associated with real news. This suggests that articles falling into this topic are more likely to be authentic. The words in real news topics include 'afp', 'photo', 'loading', 'daesh', 'eu', 'italy', 'theresa', 'march', 'david', 'european', 'brexit', 'prime', 'refugee', 'parliament', among others.

- Prevalence of Proper Nouns: Names of entities (like 'eu', 'italy', 'theresa') and organizations ('afp') are common, indicating a focus on specific, verifiable entities and locations.

- Neutral Language: The language tends to be more neutral, focusing on facts rather than emotive or sensational content.

Negative Predictors: Conversely, the model pinpointed `TextTopic7_raw14` as a cluster indicative of fake news. Articles that fit into this topic are more likely to be false or misleading. Words in fake news topics include 'police', 'protester', 'incident', 'arrest', 'authority', 'law', 'pipeline', 'shooting', 'rock', 'man', 'city', 'shot', 'spray', 'protest', 'suspect', 'activist', 'kill', 'gun', among others.

- Conflict and Violence: A significant focus on words related to conflict, law enforcement, and violence ('police', 'arrest', 'shooting', 'kill').
- Emotive Language: The presence of more emotive and potentially sensational language, possibly to evoke strong emotional responses from readers.

Conclusion:

Our text mining analysis has uncovered key characteristics that distinguish real news articles from fake ones. TextTopic7_raw6, a cluster of articles characterized by terms like 'afp,' 'eu,' and 'theresa,' is associated with real news. This cluster's prevalence of proper nouns and neutral language indicates a focus on specific, verifiable entities and factual content.

In contrast, TextTopic7_raw14, a cluster containing terms like 'police,' 'arrest,' and 'shooting,' is linked to fake news. This cluster's emphasis on conflict, violence-related terms, and emotive language suggests an attempt to manipulate readers' emotions.

Our model effectively identifies patterns and linguistic markers associated with news article authenticity. Utilizing these insights can improve automated systems for detecting and classifying real and fake news, fostering information integrity in the ever-changing digital news realm.

Business Insights:

Boosting Content Verification Tools:

The identification of specific clusters linked to real and fake news paves the way for enhanced content verification tools. A tool that harnesses the linguistic patterns identified in our study can provide businesses and media platforms with an efficient mechanism for evaluating news article authenticity.

Mitigating Risks for Platforms:

Media platforms and news aggregators can integrate our model's insights to implement risk mitigation strategies. By prioritizing articles from the positive predictor cluster and subjecting those from the negative predictor cluster to stricter scrutiny, platforms can potentially minimize the spread of fake news, bolstering their credibility.

Automated Fact-Checking Solutions:

The prevalence of proper nouns and neutral language in real news topics lays the groundwork for developing automated fact-checking solutions. Businesses can invest in systems that verify the presence of specific entities and assess language neutrality to swiftly identify and validate the authenticity of news articles before publication or sharing.

User-Facing Trust Indicators:

Implementing trust indicators for users based on our findings can enhance the user experience. Media platforms could incorporate visual cues or labels indicating the likelihood of authenticity, providing users with a quick reference to evaluate the reliability of the news they consume.

Content Moderation Strategies:

For online platforms with user-generated content, understanding the linguistic markers associated with fake news can inform content moderation strategies. By identifying content that aligns with the negative predictor cluster, platforms can implement stricter moderation measures to curb the dissemination of misleading or harmful information.

Educational Initiatives:

Businesses and media organizations can leverage our findings to develop educational initiatives aimed at improving media literacy. By educating users about the linguistic characteristics of real and fake news, individuals can become more discerning consumers of information, contributing to a more informed and resilient digital community.

Limitations:

Generalization Challenges:

The model's predictors may not generalize well across diverse contexts, impacting effectiveness in different settings.

Adaptability to Dynamic Fake News Landscape:

Staying current with evolving tactics in the dynamic fake news landscape poses challenges for the model.

Ethical Considerations in Integration:

Integrating tools into social media platforms raises ethical concerns, necessitating a delicate balance between combatting fake news and preserving freedom of expression.

References:

- Dataset Link: https://www.kaggle.com/datasets/ruchi798/source-based-news-classification
- Text Analytics Using SAS Text Miner Course Notes