Assignment 3: Topic Modelling & Sentiment Prediction

Kumar Shubham (G01402581)
Deep Vora (G01388910)

Introduction:

Topic modeling is an unsupervised machine learning technique that automatically analyzes text data to determine cluster words for a set of documents. It is capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents. This is known as 'unsupervised' machine learning because it doesn't require a predefined list of tags or training data previously classified by humans. Since topic modeling doesn't require training, it's a quick and easy way to start analyzing your data. Topic modeling techniques, such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) are both powerful tools. LDA is a good choice for applications where topic coherence and sparsity are important, while NMF is a good choice for applications where scalability and versatility are important.
Sentiment prediction, on the other hand, focuses on determining the emotional sentiment conveyed in a piece of text. This task typically involves classifying text as positive, negative, or neutral based on the presence of sentiment-bearing words and phrases. Sentiment prediction has become increasingly important in understanding customer opinions, evaluating product reviews, and gauging public sentiment toward various issues.

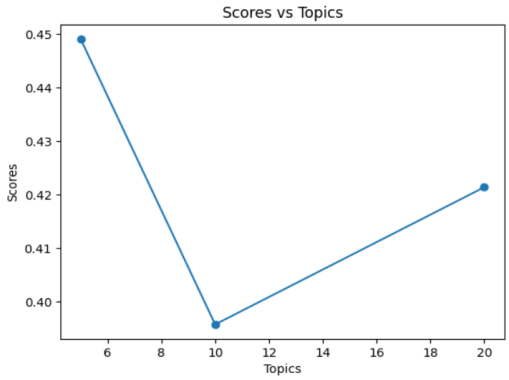
Data Pre-Processing:

WordCloud showing the cleaned data. We first pre-processed the tweet column to remove extra non-required characters, punctuations, spaces, etc.

Screenshot 1: WordCloud after data pre-processing

How did we tune our model for the number of topics?

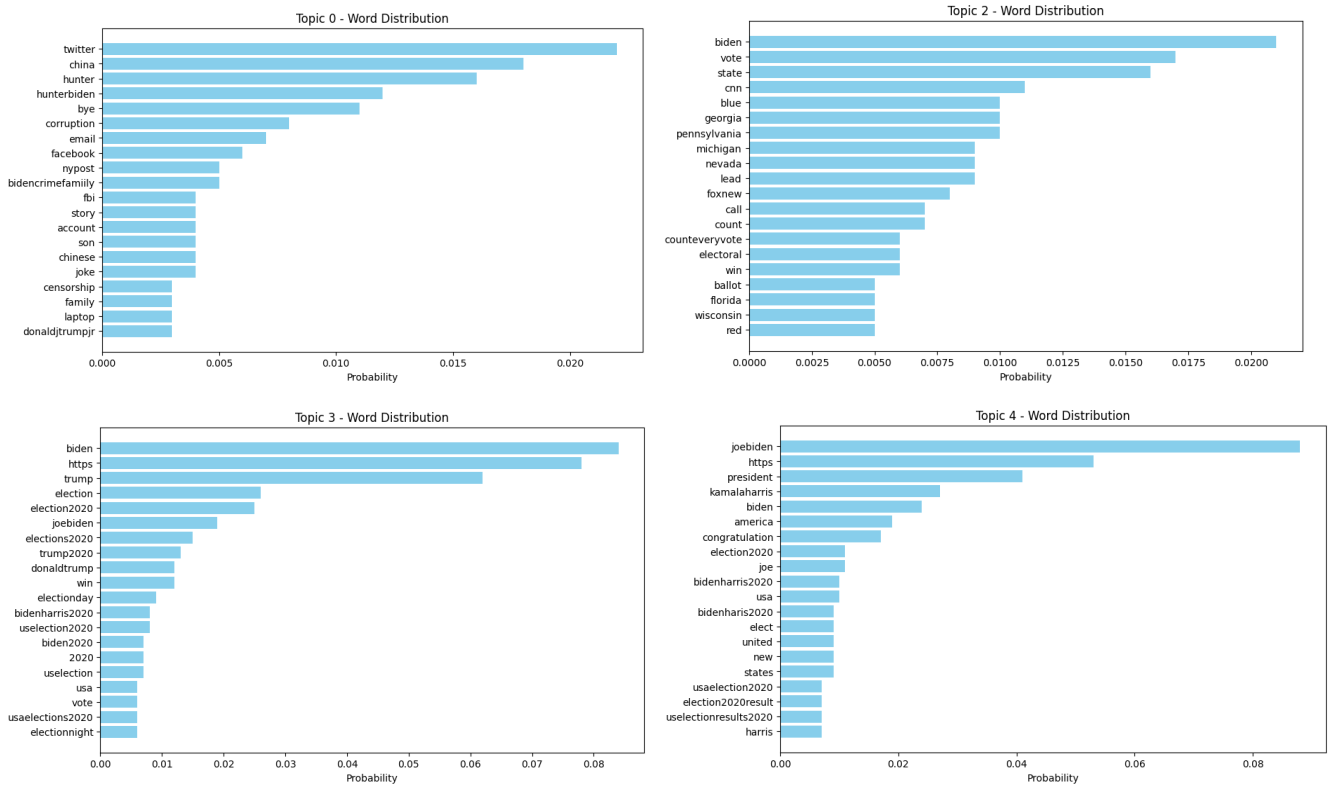
We selected random four rounds: **3, 5, 10 & 20**. And then we calculated the coherence score between these rounds. Coherence score is the cosine similarity score. It tells how close two words are with each other. We found number 5 a good number for topics because it achieved a highest **coherence score (0.45)** among others.



Screenshot 2: Score vs Topic Graph

Distribution Of Words In Each Topic:

The below graph shows the distribution of words per topic with their Topic Distribution Probabilities.



Screenshot 3: Word distribution example of Topic 0 & 2

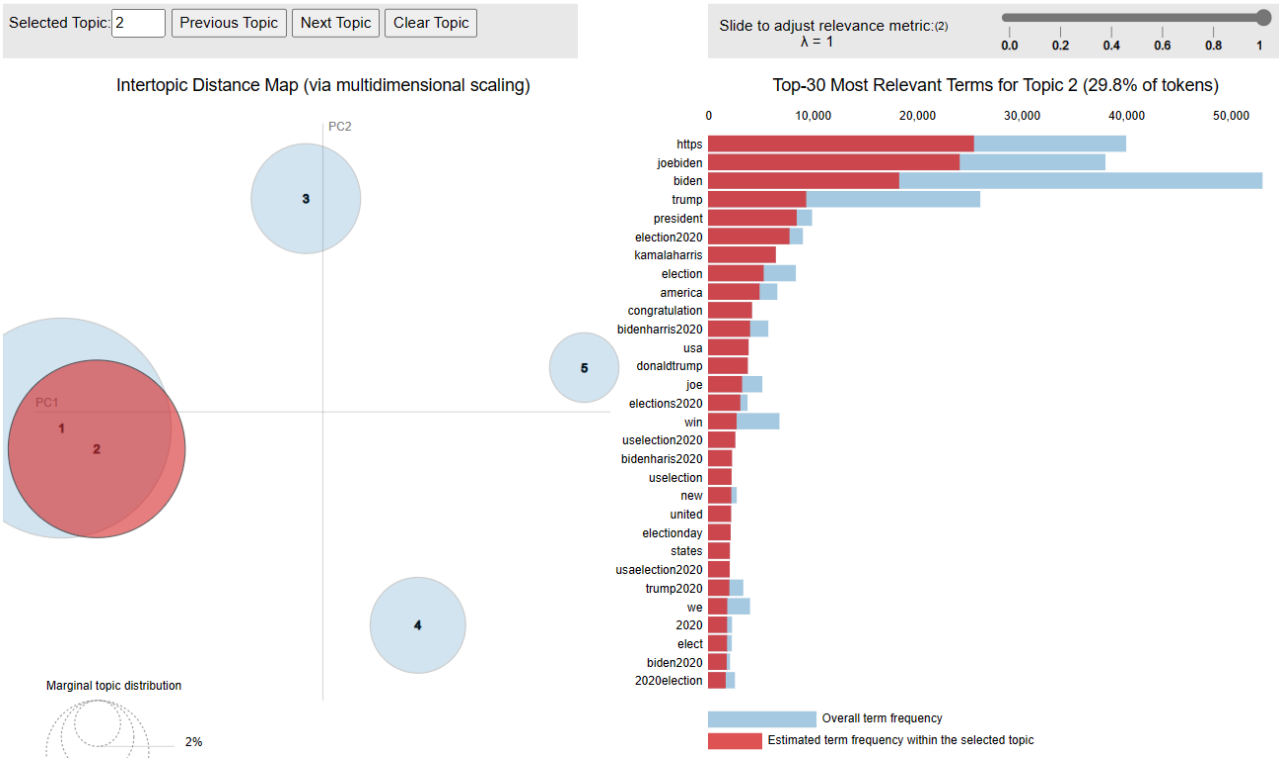
Examples of tweets that are mostly associated with each topic:

We found this by calculating the topic distribution per Document selecting the topic with the highest probability and associating the tweet with it. We have a TopTopic column to show the Topic number associated with that particular column.

index	TopTopic	tweet
1	4	us debate commission say mute democratic nominee biden us president trump microphone prevent interruption gives twominute initial response 2020uselection https tco2g80std1p8
3	4	uselections2020 joebiden continued address despite pour rain netizen reminisce obama rain rally https tco66mwenr3w6
4	4	uselection2020 survey say 65 per cent asianamerican support joebiden 28 per cent support donaldtrump https tcol8taf5yk7n
5	4	record 90 million americans vote early we presidential election datum show president donaldtrump democratic rival joebiden campaign across country try sway remain undecided voter https tco53d5btepeu
6	0	joebiden enlist alist star power help close campaign season you all think bradpitt appearance move needle https tcoxj96dlxsqp
7	4	trump vs biden whose victory help pakistan https tco8zqeuby5xn trump2020 biden uselection
8	4	uselections2020 donaldtrump joebiden hold compete rally itvideo presidentialelection https tconyk4cbqdnl
9	0	uspresidentialelection2020 get vote common theme rally donaldtrump joebiden https tcoa6rss5feyw
11	4	president donaldtrump democratic challenger joebiden look persuade early voter sunday nevada north carolina final presidential debate later week https tcofuan1cqzsz https tcomsgvr9h3ls
18	0	watch cher perform happiness thing call joe support joebiden https tcoykwhg8gycd https tcogv0tguytqr
19	4	two day electionday president donaldtrump launch campaign sprint across us battleground state start chilly outdoor rally michigan seek defy poll fend democratic challenger joebiden uspresidentialelections2020 https tcoawh3udilzq
22	4	brexitdiary biden bereitet johnson kopfschmerzen https tco1f0eplnszn brexitdiary brexit johnson biden trump
30	1	trump accuse doctor profit covid19 death biden say president give virus uselection2020 uselection donaldtrump joebiden https tcoyqhef4kut4
31	4	poll release hobby school public affairs university houston find 50 percent voter texas already vote donaldtrump 447 percent say vote joebiden uselections2020 https tconcejjeyhzu
32	1	god bette re scared fascist whitehouse racist criminal hate democracy deny science amp kill citizen electionday two day win will not country please let joebiden win thank
33	4	late news live joebiden lead donaldtrump four key us states show poll uselections2020 uselectionprediction2020 uselection https tcouemjbwcytc
36	4	president donaldtrump challenger joebiden wage pitch battle american midwest chasing every last vote four day go region propel republican victory 2016 https tcoqp9piq0pch https tcoaqxzv6zb5w
37	1	joebiden ' tax plan could affect https tco6eajokf4d1 https tcowptcs6tbtw
39	2	judicialwatch president tomfitton rudygiuliani big tech violate antitrust law censor bidenburisma scandal watch full interview https tcogrq3ddrikz https tcouyfnx0vutv
52	4	we election 2020 donald trump lead joe biden iowa via indiatvnew uselection2020 uselection donaldtrump joebiden https tcoy0d9byctll

Screenshot 4: Shows the example of tweets mostly associated with a particular Topic

pyLDAvis Visualization:



Screenshot 5: PyLDAvis Visualization for Topic 2

Interpretation of the above pyLDAvis visualization:

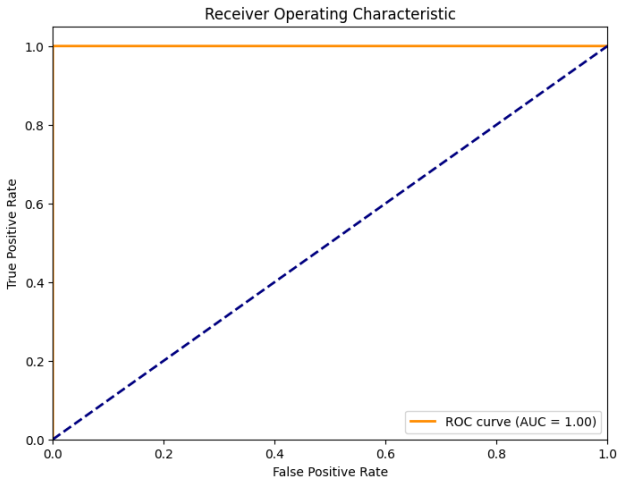
1. The pyLDAvis visualization shown above shows Topic 2, which has a 29.8% percentage of tokens.
2. The left-hand side Intertopic Distance Map shows the similarity between topics. The closer two topics are, the more similar they are. And the right-hand side is the Top-30 Most Relevant Terms for Topic 2, words that are most associated with Topic 2.

3. Topic 2 is about the 2020 US presidential election. The top 30 most relevant terms for Topic 2 include words like "https," "joebiden," "biden," "trump," "president," "election2020," "kamalaharris," "election," "america," and "congratulation bidenhams 2020."
4. The Marginal topic distribution shows that Topic 2 is the most common topic in the corpus, followed by Topic 3 and Topic 4.

ROC Curve:

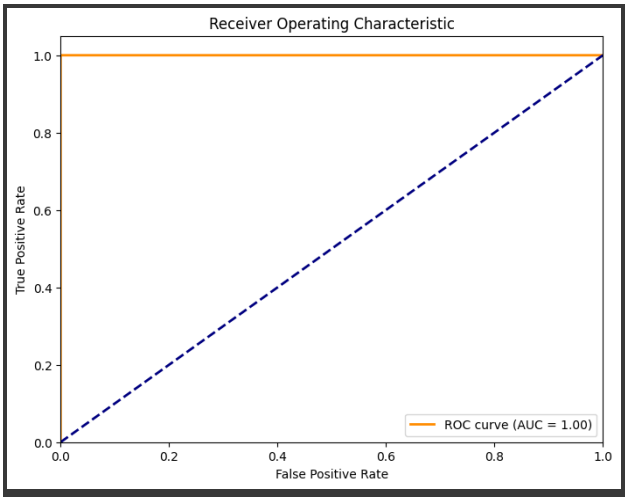
Before Topic Probability Distribution:

Logistic Regression AUC : 0.9618924156486038
Logistic Regression F1-SCORE : 0.8934267085513865



After Topic Probability Distribution:

Logistic Regression AUC : 0.9588622019932732
Logistic Regression F1-SCORE : 0.8891095179134039



Result Comparison:

We are comparing the performance of two cases for a logistic regression model, one before and one after performing Topic Probability Distribution.

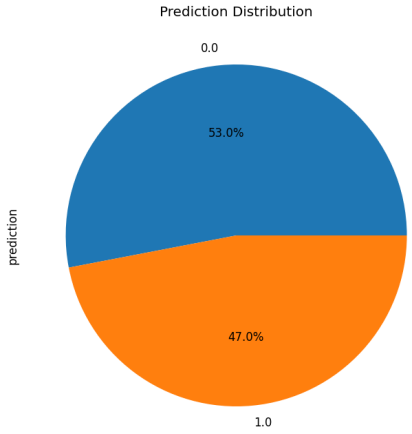
1. Area Under the Curve (AUC):

AUC is a measure of the model's ability to distinguish between the positive and negative classes. A higher AUC number denotes greater performance. The AUC achieved after applying TFIDF is marginally greater (0.9619) than the AUC achieved after enhancing tweets by adding Topic Probability Distribution (0.9589). Given that both ROC AUC values are almost 1.0 and the difference between them is negligible, it can be concluded that the model's ability to discriminate between classes is excellent in both cases.

2. F1-Score:

F1-Score represents a model's precision and recall, which tells how well the model balances false positives and false negatives The F1-score after applying TFIDF is marginally greater (0.8891) than the F1-score after performing Topic Probability Distribution (0.8934). Both the differences are very small and results indicate strong model performance.

3. Sentiment Prediction:



Conclusion:

The difference in sub-results is negligible, which indicates the model's excellent class discrimination with and without probability distribution. The F1-Scores of both the cases, which difference is also very small, supports the fact of the model accuracy. We learned about the importance of topic modeling and sentiment analysis using logistic regression.
Topic modeling helped in analyzing election tweets, identifying key topics and concerns the voters are discussing over social media. By uncovering such hidden topics and identifying sentiment, these techniques provide valuable insights that can be used to inform decision-making, improve customer engagement, and gain a deeper understanding of human language. As the volume and complexity of text data continue to grow, topic modeling and sentiment prediction will remain indispensable tools for extracting knowledge and meaning from the vast sea of information.