Bayesian_modelling

February 25, 2025

1 Airline Sentiment Analysis using Bayesian Models

1.1 Introduction

This analysis explores airline sentiment data that were scraped from Twitter. The dataset is available from the kaggle website.

The objective is to predict any outcome that might be significant from the given dataset using Bayesian modelling with PyMC.

Key areas of interest were the overall sentiment distribution, sentiment distribution per airline, factors influencing tweets from being retweeted, and user activity.

After the initial exploration, it was decided to perform a sentiment analysis using two Bayesian approaches:

- A basic model that estimates overall sentiment probabilities.
- A hierarchical model that accounts for airline-specific differences.

1.2 Data Exploration and Initial Insights

1.2.1 Dataset Overview

Load the airline sentiment dataset and examine its structure, focusing on key columns and columns with sparse data that could be cleaned.

[9]:
```
# Import required libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset
df = pd.read_csv("data/Tweets.csv")

# Show basic information about the dataset
df.info()

# Time period of the data
df['tweet_created'] = pd.to_datetime(df['tweet_created']).dt.date
print(f"Time period from {df['tweet_created'].min()} to {df['tweet_created'].
  ↪max()}")
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14640 entries, 0 to 14639
Data columns (total 15 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   tweet_id                       14640 non-null  int64
 1   airline_sentiment              14640 non-null  object
 2   airline_sentiment_confidence   14640 non-null  float64
 3   negativereason                 9178 non-null   object
 4   negativereason_confidence      10522 non-null  float64
 5   airline                        14640 non-null  object
 6   airline_sentiment_gold         40 non-null     object
 7   name                           14640 non-null  object
 8   negativereason_gold            32 non-null     object
 9   retweet_count                  14640 non-null  int64
 10  text                           14640 non-null  object
 11  tweet_coord                    1019 non-null   object
 12  tweet_created                  14640 non-null  object
 13  tweet_location                 9907 non-null   object
 14  user_timezone                  9820 non-null   object
dtypes: float64(2), int64(2), object(11)
memory usage: 1.7+ MB
Time period from 2015-02-16 to 2015-02-24
```

The *airline_sentiment_gold* and *negativereason_gold* columns do not contain much data. Therefore, drop these columns.

[10]:
```
df.pop('airline_sentiment_gold')
df.pop('negativereason_gold')

df.head()
```

[10]:

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	name	retweet_count	text
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	cairdin	0	@VirginAmerica What @dhepburn said.
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	jnardino		
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	yvonnalynn		
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	jnardino		
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	jnardino		

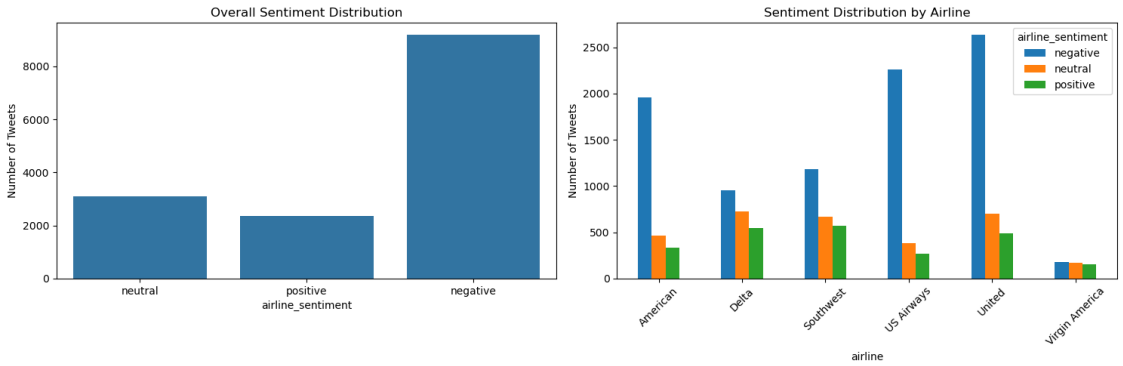
```
1              0  @VirginAmerica plus you've added commercials t…
2              0  @VirginAmerica I didn't today… Must mean I n…
3              0  @VirginAmerica it's really aggressive to blast…
4              0  @VirginAmerica and it's a really big bad thing…
```

```
  tweet_coord tweet_created tweet_location               user_timezone
0         NaN    2015-02-24            NaN  Eastern Time (US & Canada)
1         NaN    2015-02-24            NaN  Pacific Time (US & Canada)
2         NaN    2015-02-24      Lets Play  Central Time (US & Canada)
3         NaN    2015-02-24            NaN  Pacific Time (US & Canada)
4         NaN    2015-02-24            NaN  Pacific Time (US & Canada)
```

1.2.2 Sentiment Distributions

Inspect the overall sentiment distribution as well as the sentiment distribution by airline.

[11]:
```
# Create sentiment distribution plots
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(15, 5))

# Overall sentiment distribution
sns.countplot(data=df, x='airline_sentiment', ax=ax1)
ax1.set_title('Overall Sentiment Distribution')
ax1.set_ylabel('Number of Tweets')

# Sentiment distribution by airline
sentiment_by_airline = pd.crosstab(df['airline'], df['airline_sentiment'])
sentiment_by_airline.plot(kind='bar', ax=ax2)
ax2.set_title('Sentiment Distribution by Airline')
ax2.set_ylabel('Number of Tweets')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()

# Print percentage distribution
print("\nSentiment distribution (%):")
print(df['airline_sentiment'].value_counts(normalize=True).round(3) * 100)
```

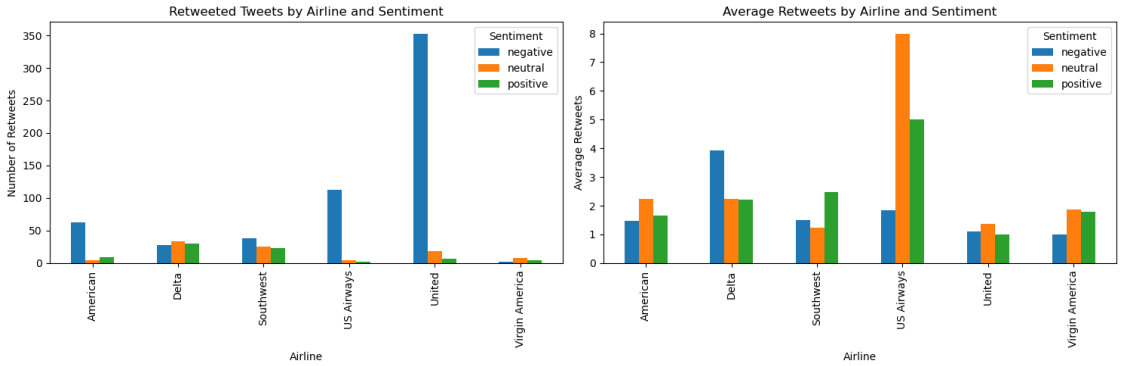


```
Sentiment distribution (%):
airline_sentiment
negative    62.7
neutral     21.2
positive    16.1
Name: proportion, dtype: float64
```

1.2.3 Engagement Analysis

Examining tweet engagement through retweets provides insight into which types of feedback gain more traction.

[12]:
```
# Analyze retweeted tweets
df_retweeted = df[df['retweet_count'] > 0]

fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(15, 5))

# Number of retweeted tweets by sentiment and airline
retweeted_counts = df_retweeted.groupby(['airline', 'airline_sentiment']).
 ↪size().unstack()
retweeted_counts.plot(kind='bar', ax=ax1)
ax1.set_title('Retweeted Tweets by Airline and Sentiment')
ax1.set_xlabel('Airline')
ax1.set_ylabel('Number of Retweets')
ax1.legend(title='Sentiment')

# Average retweets when retweeted
avg_retweets = df_retweeted.groupby(['airline',␣
 ↪'airline_sentiment'])['retweet_count'].mean().unstack()
avg_retweets.plot(kind='bar', ax=ax2)
ax2.set_title('Average Retweets by Airline and Sentiment')
ax2.set_xlabel('Airline')
ax2.set_ylabel('Average Retweets')
ax2.legend(title='Sentiment')

plt.tight_layout()
plt.show()
```



The following observations are seen from the above exploration:

- Generally, negative tweets seem to be retweeded the most, especially for the United airline.
- The retweet distribution varies across airlines.
- The average number of neatural and positive sentiment retweets for the US Airways airline is significatly higher than seen for the other airlines.

1.2.4 User Activity

Inspect user tweeting patterns for possible signs of spammers or bots that might skew the data.

[13]:
```
# Analyze user tweeting patterns
user_tweet_counts = df['name'].value_counts()

print("Top users by number of tweets:")
print(user_tweet_counts.head(10))

# Get the most active user's name and tweets
most_active_user = df['name'].value_counts().index[0]
most_active_user_tweets = df[df['name'] == most_active_user]

# Visually inspect tweets from most active user
print(f"Sample tweets from user: {most_active_user}")
print("-" * 80)

sample_tweets = df[df['name'] == most_active_user][['tweet_created', 'airline',␣
 ↪'airline_sentiment', 'text']].head(5)

# Format and display each tweet
for idx, tweet in sample_tweets.iterrows():
    print(f"Date: {tweet['tweet_created']}")
    print(f"Airline: {tweet['airline']}")
    print(f"Sentiment: {tweet['airline_sentiment']}")
    print(f"Tweet: {tweet['text']}")
```

```
    print("-" * 80)

# Calculate sentiment distribution
sentiment_distribution = most_active_user_tweets['airline_sentiment'].
 ↪value_counts()
sentiment_percentages = (sentiment_distribution / len(most_active_user_tweets)␣
 ↪* 100).round(2)

print(f"\nSentiment distribution for {most_active_user}'s tweets:")
print(f"Total tweets: {len(most_active_user_tweets)}")
for sentiment, percentage in sentiment_percentages.items():
    print(f"{sentiment}: {percentage}% ({sentiment_distribution[sentiment]}␣
 ↪tweets)")

# Calculate the number of tweets per airline for the most active user
airline_tweet_counts = most_active_user_tweets['airline'].value_counts()

print(f"\nNumber of tweets per airline for {most_active_user}:")
for airline, count in airline_tweet_counts.items():
    print(f"{airline}: {count} tweets")
```

```
Top users by number of tweets:
name
JetBlueNews        63
kbosspotter        32
_mhertz            29
otisday            28
throthra           27
weezerandburnie    23
rossj987           23
GREATNESSEOA       22
MeeestarCoke       22
scoobydoo9749      21
Name: count, dtype: int64
Sample tweets from user: JetBlueNews
--------------------------------------------------------------------------------
Date: 2015-02-20
Airline: Virgin America
Sentiment: neutral
Tweet: @VirginAmerica achieves a second year of profitability despite revenue
pressure … - @CAPA_Aviation http://t.co/zSuZTNAIJq
--------------------------------------------------------------------------------
Date: 2015-02-21
Airline: Delta
Sentiment: neutral
Tweet: @JetBlue Fliers to Gain Access to WSJ Content - Analyst Blog - Nasdaq
http://t.co/dWEse7Xidr
```

```
--------------------------------------------------------------------------------
Date: 2015-02-21
Airline: Delta
Sentiment: neutral
Tweet: @JetBlue Airways Now Covered by Bank of America (JBLU) - Dakota Financial
News http://t.co/WOwxhpU5qv
--------------------------------------------------------------------------------
Date: 2015-02-21
Airline: Delta
Sentiment: neutral
Tweet: @JetBlue Airways Stock Rating Lowered by Vetr Inc. (JBLU) - Dakota
Financial News http://t.co/QW2eBEEMVg
--------------------------------------------------------------------------------
Date: 2015-02-21
Airline: Delta
Sentiment: neutral
Tweet: @JetBlue's CEO battles to appease passengers and Wall Street - Waterbury
Republican American http://t.co/FFc13zYGJS
--------------------------------------------------------------------------------

Sentiment distribution for JetBlueNews's tweets:
Total tweets: 63
neutral: 90.48% (57 tweets)
positive: 7.94% (5 tweets)
negative: 1.59% (1 tweets)

Number of tweets per airline for JetBlueNews:
Delta: 62 tweets
Virgin America: 1 tweets
```

- The top tweeter is JetBlueNews, which seems to be a news bot.
- The sentiment of JetBlueNews' tweets are mostly neutral.
- JetBlueNews mostly tweets for the Delta airline.
- The tweets from the other high-engagement users seem balanced and normal. In other words, it does not seem like any users are disproportionarly influencing sentiment trends.

1.2.5 Findings

The following observations were made from the exploratory analysis:

- There are significantly more tweets of negative sentiment than of neutral or positive sentiment.
- Sentiment distribution varies across airlines.
- Negative sentiments tend to get more engagement through retweets.
- The user with the highest number of tweets seems to be a bot that shares news with a predominantly neutral sentiment.

1.3 Modeling Approach

Based on our exploratory analysis, it was decided to focus on modeling airline sentiment using a Bayesian approach.

Bayesian modelling is a statistical approach that incorporates prior knowledge along with observed data to estimate the probability of different outcomes. This approach is particularly useful when dealing with uncertainty and limited data.

1.3.1 Bayesian Modelling Overview

Bayesian modelling allows:

- Starting with prior assumptions about sentiment probabilities.
- Updating these assumptions using observed data.
- Quantifying uncertainty in the predictions.

1.3.2 Model Implementation

Two models of increasing complexity will be implemented:

1. **Basic Model**
    - Predicts overall sentiment probabilities across all airlines
    - Uses a Dirichlet prior (assuming equal probabilities for all sentiments)
    - Estimates probability of a tweet being negative, neutral, or positive
2. **Hierarchical Model**
    - Extends the basic model to account for airline-specific patterns
    - Allows sentiment probabilities to vary by airline
    - Shares information across airlines while preserving airline-specific patterns

1.3.3 Data Preparation

To implement these models, the following preparation will be done:

- Map sentiment categories to numerical values (negative=0, neutral=1, positive=2)
- Create airline-specific datasets for the hierarchical model
- Clean and format the data for PyMC modeling

[14]:
```
# Clean the dataset
df = df[['airline_sentiment',
         'airline',
         'text',
         'tweet_created']]

# Map sentiment categories to numerical values (required for PyMC's categorical␣
 ↪distribution)
```

```
sentiment_map = {'negative': 0, 'neutral': 1, 'positive': 2}
df['sentiment_categories'] = df['airline_sentiment'].map(sentiment_map)

# Create airline-specific datasets
airlines = df['airline'].unique()
airline_data = {
    airline: {
        'counts': df[df['airline'] == airline]['sentiment_categories'].
 ↪value_counts().reindex([0, 1, 2], fill_value=0).values,
        'total': len(df[df['airline'] == airline])
    }
    for airline in airlines
}

# Display the first few rows of the cleaned dataset
df.head()
```

[14]:

	airline_sentiment	airline	text	tweet_created	sentiment_categories
0	neutral	Virgin America	@VirginAmerica What @dhepburn said.	2015-02-24	1
1	positive	Virgin America	@VirginAmerica plus you've added commercials t…	2015-02-24	2
2	neutral	Virgin America	@VirginAmerica I didn't today… Must mean I n…	2015-02-24	1
3	negative	Virgin America	@VirginAmerica it's really aggressive to blast…	2015-02-24	0
4	negative	Virgin America	@VirginAmerica and it's a really big bad thing…	2015-02-24	0

1.3.4 Basic Bayesian Model

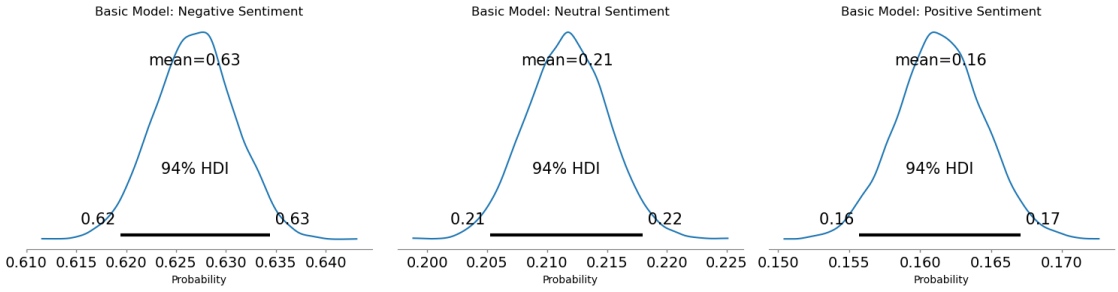
Implement a simple Dirichlet-Categorical model to estimate overall sentiment probabilities across all airlines. Include prior selection and posterior analysis.

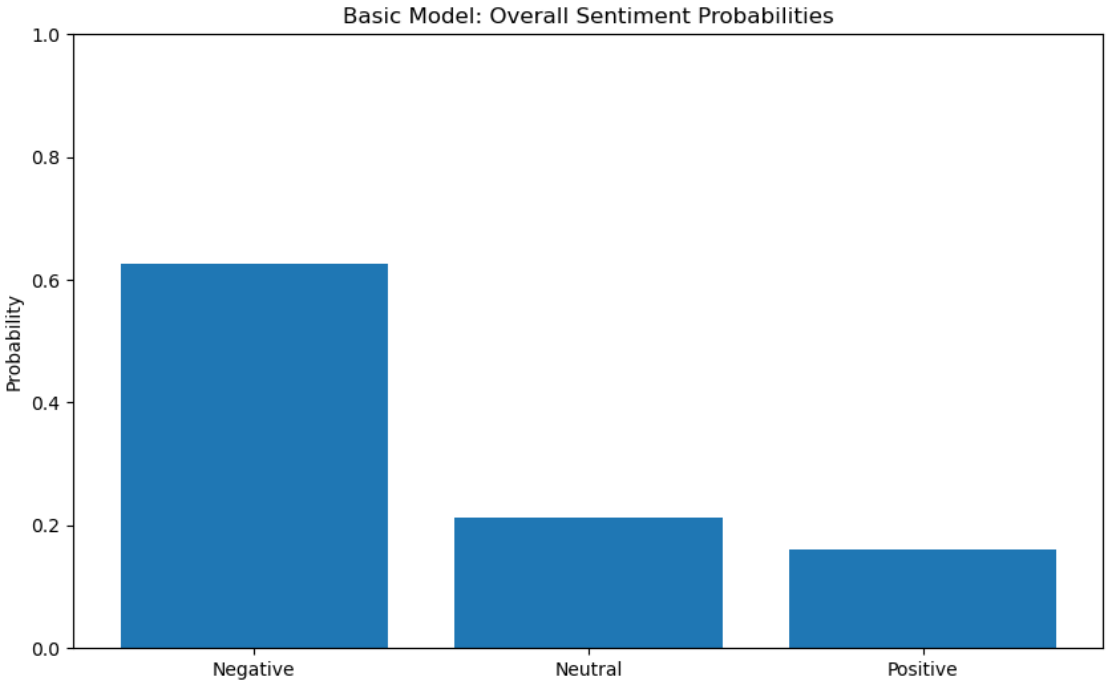
[15]:
```
import pymc as pm
import numpy as np
import arviz as az
```

```
# Count the number of tweets in each sentiment category
sentiment_counts = df['airline_sentiment'].value_counts().sort_index().values
print(f"Number of tweets per sentiment category (negative, neutral, positive):␣
 ↪{sentiment_counts}")

# Set a prior: symmetric Dirichlet prior
alpha_prior = np.array([1, 1, 1])  # Weak prior, assumes no prior knowledge

# Define the model
with pm.Model() as basic_model:
    # Dirichlet prior for probabilities of each sentiment
    theta = pm.Dirichlet("theta", a=alpha_prior)

    # Observed categorical data
    observed = pm.Categorical("observed", p=theta, observed=np.repeat([0, 1,␣
 ↪2], sentiment_counts))

    # Sample from the posterior
    basic_trace = pm.sample(2000, return_inferencedata=True)

# Summary of posterior distributions
az.summary(basic_trace, var_names=["theta"])

# # Plot posterior distributions
# az.plot_posterior(basic_trace, var_names=["theta"])
# plt.show()

# Create meaningful labels
sentiment_labels = ['Negative', 'Neutral', 'Positive']
airline_labels = df['airline'].unique()

# Basic Model Plots
fig, axs = plt.subplots(1, 3, figsize=(15, 4))
for i, sentiment in enumerate(sentiment_labels):
    az.plot_posterior(basic_trace,
                      var_names="theta",
                      coords={'theta_dim_0': i},
                      ax=axs[i])
    axs[i].set_title(f"Basic Model: {sentiment} Sentiment")
    axs[i].set_xlabel("Probability")

plt.tight_layout()
plt.show()

def plot_basic_sentiment_probabilities(basic_trace):
    plt.figure(figsize=(10, 6))
```

```
    # Extract posterior means
    posterior_means = az.summary(basic_trace, var_names=['theta'])

    # Create bar plot
    sentiment_labels = ['Negative', 'Neutral', 'Positive']
    values = [posterior_means.loc[f'theta[{i}]', 'mean'] for i in range(3)]

    plt.bar(sentiment_labels, values)
    plt.ylim(0, 1)
    plt.title('Basic Model: Overall Sentiment Probabilities')
    plt.ylabel('Probability')
    plt.show()

# Plot the basic model results
plot_basic_sentiment_probabilities(basic_trace)
```

```
Initializing NUTS using jitter+adapt_diag…

Number of tweets per sentiment category (negative, neutral, positive): [9178
3099 2363]

Multiprocess sampling (4 chains in 4 jobs)
NUTS: [theta]

Output()


Sampling 4 chains for 1_000 tune and 2_000 draw iterations (4_000 + 8_000 draws
total) took 8 seconds.
```





Basic Model Evaluation The model estimates the following probabilities for each sentiment:

Negative Sentiment - Mean probability: 0.63 (63% of tweets are negative) - 94% HDI: [0.62, 0.63]

Neutral Sentiment - Mean probability: 0.21 (21% of tweets are neutral) - 94% HDI: [0.21, 0.22]

Positive Sentiment - Mean probability: 0.16 (16% of tweets are positive) - 94% HDI: [0.16, 0.17]

The Highest Density Interval (HDI) represents the most credible values of the parameter estimate: - 94% of the posterior probability falls within this interval - The narrow HDI intervals suggest that the model is highly confident in its estimates

1.3.5 Hierarchical Bayesian Model

Implement a hierarchical model to account for airline-specific sentiment distributions, using a population-level prior and airline-specific parameters.

[17]:
```
# Define the hierarchical model
with pm.Model() as hierarchical_model:
    # Hyperpriors for the population-level sentiment distribution
    alpha_population = pm.Dirichlet('alpha_population', a=np.ones(3))

    # Concentration parameter for airline-specific distributions
    concentration = pm.Gamma('concentration', alpha=2, beta=1)

    # Airline-specific sentiment distributions
```

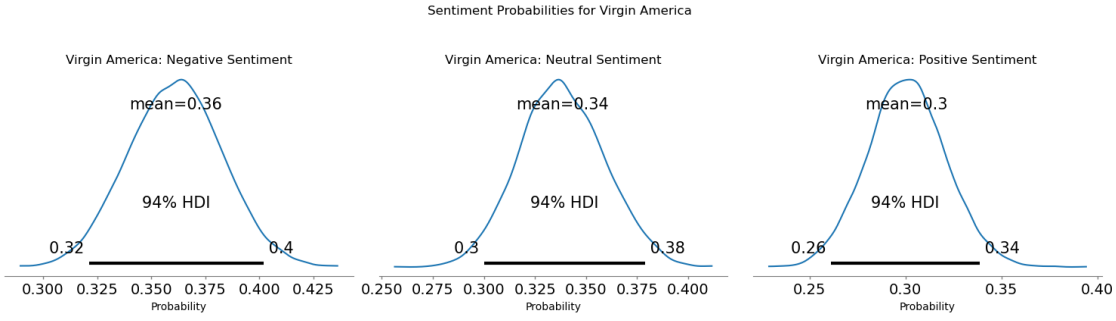
```
    theta = pm.Dirichlet('theta',
                         a=(alpha_population * concentration)[None, :],
                         shape=(len(airlines), 3))

    # Observations for each airline
    for i, airline in enumerate(airlines):
        pm.Categorical(f'obs_{airline}',
                       p=theta[i],
                       observed=np.repeat([0, 1, 2],
                                          airline_data[airline]['counts']))

    # Sample from posterior
    hierarchical_trace = pm.sample(2000, tune=1000, return_inferencedata=True)

# # Plot results
# az.plot_posterior(hierarchical_trace, var_names=['theta'])
# plt.show()

# Create meaningful labels
sentiment_labels = ['Negative', 'Neutral', 'Positive']
airline_labels = df['airline'].unique()

# Hierarchical Model Plots - one plot per airline
for airline_idx, airline in enumerate(airline_labels):
    fig, axs = plt.subplots(1, 3, figsize=(15, 4))

    for sent_idx, sentiment in enumerate(sentiment_labels):
        az.plot_posterior(hierarchical_trace,
                          var_names="theta",
                          coords={'theta_dim_0': airline_idx,
                                  'theta_dim_1': sent_idx},
                          ax=axs[sent_idx])
        axs[sent_idx].set_title(f"{airline}: {sentiment} Sentiment")
        axs[sent_idx].set_xlabel("Probability")

    plt.suptitle(f"Sentiment Probabilities for {airline}", y=1.05)
    plt.tight_layout()
    plt.show()


# Add visualization of differences between airlines
def plot_sentiment_comparisons(hierarchical_trace, airlines):
    plt.figure(figsize=(12, 8))

    # Extract posterior means for each airline
    posterior_means = az.summary(hierarchical_trace, var_names=['theta'])
```

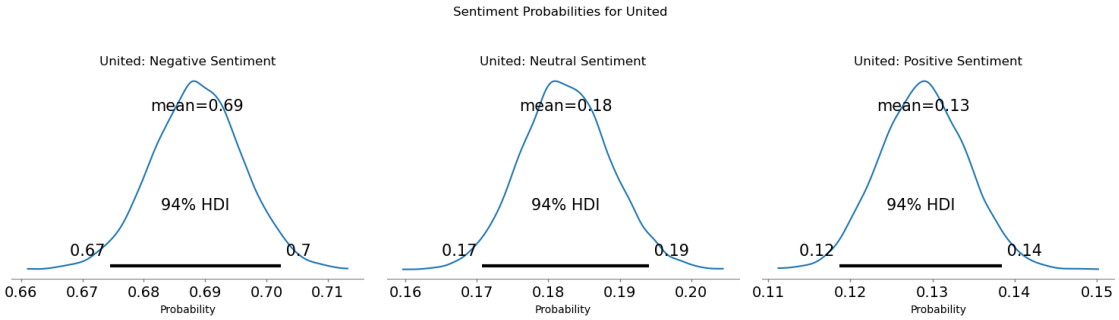
```
    # Create sentiment comparison plot
    sentiment_labels = ['Negative', 'Neutral', 'Positive']
    x = np.arange(len(airlines))
    width = 0.25

    for i, sentiment in enumerate(sentiment_labels):
        values = []
        for j in range(len(airlines)):
            key = f'theta[{j}, {i}]'
            values.append(posterior_means.loc[key, 'mean'])

        plt.bar(x + i*width, values, width, label=sentiment)

    plt.xlabel('Airlines')
    plt.ylabel('Probability')
    plt.ylim(0, 1)
    plt.title('Hierarchical Model: Posterior Mean Sentiment Probabilities by␣
 ↪Airline')
    plt.xticks(x + width, airlines, rotation=45)
    plt.legend()
    plt.tight_layout()
    plt.show()

# Bar charts comparing sentiment probabilities between airlines
plot_sentiment_comparisons(hierarchical_trace, airlines)
```

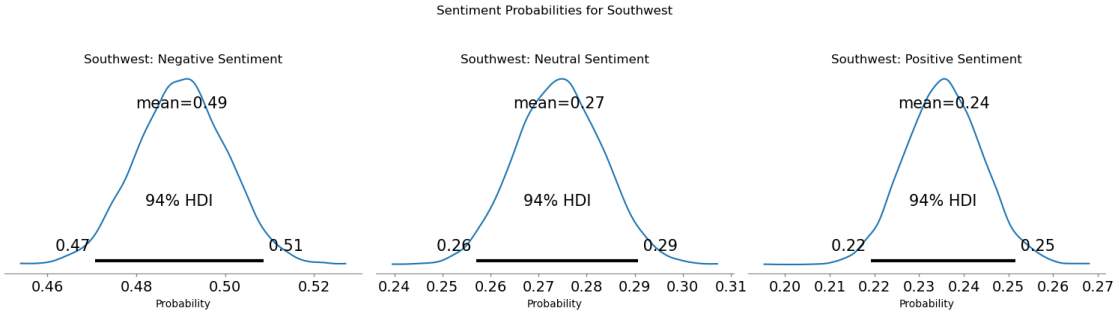
```
Initializing NUTS using jitter+adapt_diag…
Multiprocess sampling (4 chains in 4 jobs)
NUTS: [alpha_population, concentration, theta]
```

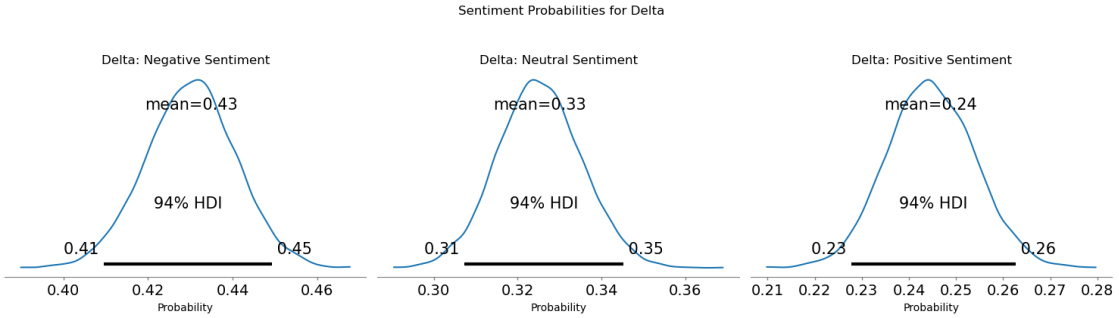
```
Output()
```

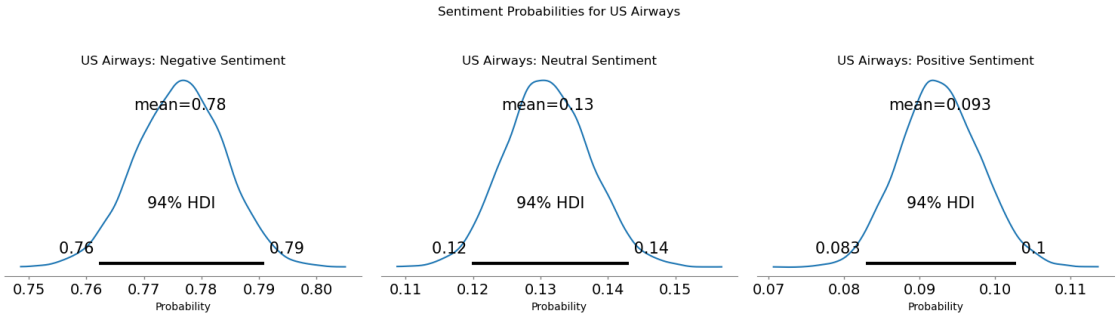
```
Sampling 4 chains for 1_000 tune and 2_000 draw iterations (4_000 + 8_000 draws
total) took 29 seconds.
```

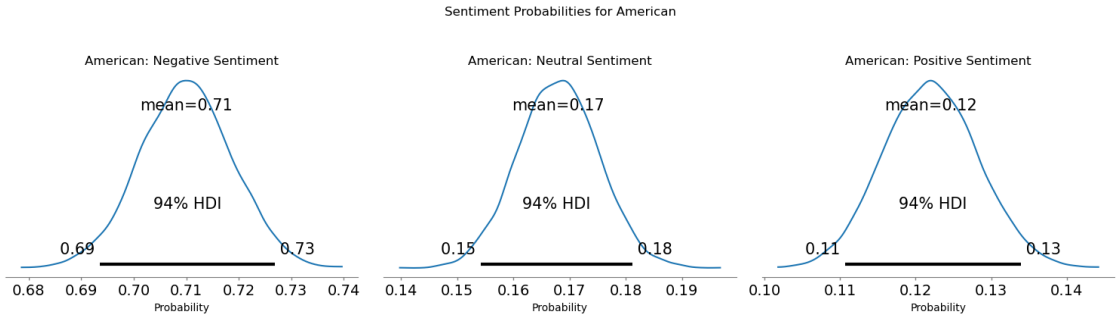


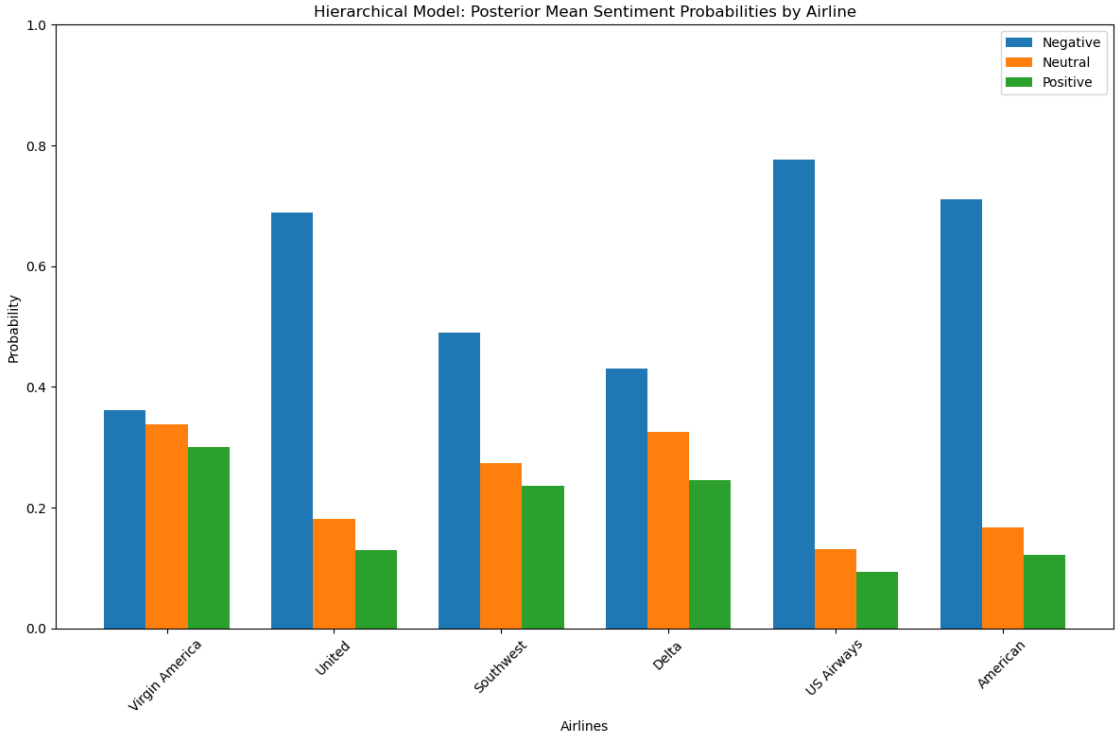

Sentiment Probabilities for United
United: Negative Sentiment
mean=0.69
94% HDI
0.67
0.7
Probability
United: Neutral Sentiment
mean=0.18
94% HDI
0.17
0.19
Probability
United: Positive Sentiment
mean=0.13
94% HDI
0.12
0.14
Probability


Sentiment Probabilities for Southwest
Southwest: Negative Sentiment
mean=0.49
94% HDI
0.47
0.51
Probability
Southwest: Neutral Sentiment
mean=0.27
94% HDI
0.26
0.29
Probability
Southwest: Positive Sentiment
mean=0.24
94% HDI
0.22
0.25
Probability


Sentiment Probabilities for Delta
Delta: Negative Sentiment
mean=0.43
94% HDI
0.41
0.45
Probability
Delta: Neutral Sentiment
mean=0.33
94% HDI
0.31
0.35
Probability
Delta: Positive Sentiment
mean=0.24
94% HDI
0.23
0.26
Probability


Sentiment Probabilities for US Airways
US Airways: Negative Sentiment
mean=0.78
94% HDI
0.76
0.79
Probability
US Airways: Neutral Sentiment
mean=0.13
94% HDI
0.12
0.14
Probability
US Airways: Positive Sentiment
mean=0.093
94% HDI
0.083
0.1
Probability





Hierarchical Model Evaluation The model's HDI intervals are generally narrow, indicating high degrees of confidence in estimates.

The HDI interval for the Virgin America airline is slightly wider. It was seen earlier that the number of tweets for this airline is comparatively small compared to the other airlines, which could cause the wider interval. There might also be more variability or weaker patterns in that data.

Furthermore, the following observations can be drawn from the hierarchical model analysis:

- For all the examined airlines, the chances of people posting negative tweets are higher than for positive or negative tweets. Although negative news is most likely to travel, this is a troubling finding.

- More specifically, the probability that people will post negative tweets about the United, US Airways and American airways is above 60%.
- The probability that tweets will be negative, neutral or positive is well balanced for the Virgin America airline.

1.4 Conclusion

The key insight from the sentiment analysis is that there is a staggering 63% chance that future tweets about an airline will be negative.

The probability for negative tweets about the United, US Airways and American airways are particularly high, all being above 60%.

It is recommended that the airlines conduct a deeper analysis into the reasons for the negative sentiment. Natural language processing techniques can be used to identify the most common negative words and phrases in the tweets. This can help the airlines to understand the root causes of the negative sentiment and develop strategies to address them.