In [ ]:
```
!pip install openpyxl
```

In [ ]:
```
!pip install wordcloud
```

In [ ]:
```
!pip install gensim
```

Purpose:

This notebook aims to analyze three datasets from Youtube, Facebook, and Reddit, performing preprocessing and LDA topic modeling for each label and overall, as specified in the requirements.

Project Requirements:

1. Analyze three CSV datasets:

- YouTube data
- Facebook comments dataset
- Reddit comments dataset

2. Perform LDA topic modeling for each label:

- Labels: 1. Cash, 2. CPF MediSave Account Top-Ups, 3. Personal Income Tax Rebate, 4. CPF Retirement or Special Account, 5. NS LifeSG credits, 6. CDC voucher, 7. U-Save rebates, 8. S&CC rebates, 9. Education/Training, 10. Retrenchment, and 11. General
- Total: 12 Topic models (11 labels and 1 overall)

3. Document each section with purpose, goals, and analytical insights.

Step 1: Data Loading

Purpose:

To load and inspect the datasets to understand their structure and contents.

Goal:

Ensure the data is correctly loaded and identify any immediate issues or preprocessing needs.

Loading the Datasets

Let's load and inspect the three datasets: YouTube, Facebook, and Reddit.

In [10]:
```
import pandas as pd

# Load YouTube dataset
youtube_data = pd.read_csv('youtube_data.csv')
print("YouTube Data:")
print(youtube_data.head())
print(youtube_data.info())

# Load Facebook dataset
facebook_data = pd.read_csv('Compiled_Facebook_Mothership_Comments_Dataset_with_Cat
print("\nFacebook Data:")
print(facebook_data.head())
print(facebook_data.info())

# Load Reddit dataset
reddit_data = pd.read_excel('reddit comment with categories.xlsx')
print("\nReddit Data:")
print(reddit_data.head())
print(reddit_data.info())
```

YouTube Data:

	profileName	text	\
0	@2011crane	CPF is to ensure a self funded pension If RA i...	
1	@77naaz	🎉yay !!	
2	@ArabicReja973	After announcing his budget, the first thing M...	
3	@ArabicReja973	After announcing his budget, the first thing M...	
4	@AutonomousSystem---19	The point is why doesnt it included together a...	

	date	Social Media	Cash	CPF Medisave	Income Tax Rebate	\
0	2024-02-17T02:48:30Z	YouTube	0	0	0	
1	2024-02-17T10:23:30Z	YouTube	0	0	0	
2	2024-02-16T14:41:21Z	YouTube	1	0	0	
3	2024-02-19T15:23:04Z	YouTube	1	0	0	
4	2024-03-01T13:33:50Z	YouTube	0	0	0	

	CPF RA -SA	NS Credits	CDC Vouchers	U-Save Rebates	S&CC Rebates	\
0	1	0	0	0	0	
1	0	0	0	0	0	
2	0	0	1	0	0	
3	0	0	0	0	0	
4	0	0	0	0	0	

	Education/Training	Retrenchment	General
0	0	0	0
1	1	1	0
2	0	0	0
3	1	1	0
4	1	0	0

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 556 entries, 0 to 555
Data columns (total 15 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   profileName         556 non-null    object
 1   text                556 non-null    object
 2   date                556 non-null    object
 3   Social Media        556 non-null    object
 4   Cash                556 non-null    int64
 5   CPF Medisave        556 non-null    int64
 6   Income Tax Rebate   556 non-null    int64
 7   CPF RA -SA          556 non-null    int64
 8   NS Credits          556 non-null    int64
 9   CDC Vouchers        556 non-null    int64
 10  U-Save Rebates      556 non-null    int64
 11  S&CC Rebates        556 non-null    int64
 12  Education/Training  556 non-null    int64
 13  Retrenchment        556 non-null    int64
 14  General             556 non-null    int64
dtypes: int64(11), object(4)
memory usage: 65.3+ KB
None
```

Facebook Data:

	postTitle	\
0	2024 saw a Budget with very little in the way ...	
1	2024 saw a Budget with very little in the way ...	

```
2  2024 saw a Budget with very little in the way ...
3  2024 saw a Budget with very little in the way ...
4  2024 saw a Budget with very little in the way ...

                                          profileId             profileName  \
0  pfbid02ju9uRcX2XTnDsvBToKNTANhzv37Byma4QNf3m8P...  Derrick Jason De Costa
1  pfbid02aXWJPVoqGLKrhe9r6X8WKRFKHrapdkNJciFVngJ...       Arulsamy Anthony
2  pfbid0Jp3EyFgxXRhuJCwve1V74bNtDp6sGJ2HUrzuD6rY...               Abu Rabi
3  pfbid0HKYZnjwTFkG23FJh43cvyyaT21xDiuNAuSfjHdWk...         Saqib Mohammad
4                                        6.16E+13        Affected Citizen

                                               text             date  \
0  I often wonder, when the PAP criticizes opposi...  17/2/2024 13:32
1  Where in the world  can you get a future PM li...  17/2/2024 11:48
2  Thank u sir!!!!😀👍\nU r the best person & will ...  17/2/2024 19:05
3  Lawrence should sing the budget in parliament ...  17/2/2024 18:43
4  So is the ministers pay increasing or still th...  17/2/2024 23:20

  Social Media  Cash  CPF Medisave  Income Tax Rebate  CPF RA -SA  NS Credits  \
0     Facebook     1             1                  1           0           1
1     Facebook     1             1                  1           0           1
2     Facebook     1             1                  1           0           1
3     Facebook     1             1                  1           0           1
4     Facebook     1             1                  1           0           1

   CDC Vouchers  U-Save Rebates  S&CC Rebates  Education/Training  \
0             1               1             1                   0
1             1               1             1                   0
2             1               1             1                   0
3             1               1             1                   0
4             1               1             1                   0

   Retrenchment  General
0             0        0
1             0        0
2             0        0
3             0        0
4             0        0
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1309 entries, 0 to 1308
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   postTitle           1309 non-null   object
 1   profileId           1309 non-null   object
 2   profileName         1309 non-null   object
 3   text                1270 non-null   object
 4   date                1309 non-null   object
 5   Social Media        1309 non-null   object
 6   Cash                1309 non-null   int64
 7   CPF Medisave        1309 non-null   int64
 8   Income Tax Rebate   1309 non-null   int64
 9   CPF RA -SA          1309 non-null   int64
 10  NS Credits          1309 non-null   int64
 11  CDC Vouchers        1309 non-null   int64
 12  U-Save Rebates      1309 non-null   int64
```

```
 13  S&CC Rebates        1309 non-null   int64
 14  Education/Training  1309 non-null   int64
 15  Retrenchment        1309 non-null   int64
 16  General             1309 non-null   int64
dtypes: int64(11), object(6)
memory usage: 174.0+ KB
None

Reddit Data:
                                            Comments  \
0  deputy prime minister lawrence wong will be de...
1  wp has spoken on the need for a proactive ai i...
2  oh shit i thought this was going a whole diffe...
3  thanks for looking into it. i am not intereste...
4  [ite progression award](https://i.imgur.com/ld...

                  createdAt              username Platform   Cash  \
0  2024-02-16T01:01:23.000Z             KeythKatz    Reddit   0.0
1  2024-02-16T16:06:52.000Z   Ambitious_Map_Kiddo    Reddit   0.0
2  2024-02-17T01:14:46.000Z         redditor_here    Reddit   0.0
3  2024-02-17T02:24:59.000Z  Professional_Race351    Reddit   0.0
4  2024-02-16T08:16:58.000Z             KeythKatz    Reddit   0.0

   CPF MediSave  Personal Income Tax Rebate  CPF Retirement/Special Account  \
0           0.0                         0.0                             0.0
1           0.0                         0.0                             0.0
2           0.0                         0.0                             0.0
3           0.0                         0.0                             0.0
4           1.0                         0.0                             0.0

   NS LifeSG credits  CDC Vouchers  U-save rebates  S&CC rebates  Education  \
0                0.0           0.0             0.0           0.0        0.0
1                0.0           0.0             0.0           0.0        0.0
2                0.0           0.0             0.0           0.0        0.0
3                0.0           0.0             0.0           0.0        0.0
4                0.0           0.0             0.0           0.0        1.0

   Retrenchment
0           0.0
1           0.0
2           0.0
3           0.0
4           0.0
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 14 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   Comments                       887 non-null    object
 1   createdAt                      890 non-null    object
 2   username                       890 non-null    object
 3   Platform                       891 non-null    object
 4   Cash                           890 non-null    float64
 5   CPF MediSave                   890 non-null    float64
 6   Personal Income Tax Rebate     890 non-null    float64
 7   CPF Retirement/Special Account  890 non-null    float64
```

```
 8   NS LifeSG credits              890 non-null    float64
 9   CDC Vouchers                   890 non-null    float64
 10  U-save rebates                 890 non-null    float64
 11  S&CC rebates                   890 non-null    float64
 12  Education                      890 non-null    float64
 13  Retrenchment                   890 non-null    float64
dtypes: float64(10), object(4)
memory usage: 97.6+ KB
None
```

Data Inspection Report

YouTube Data Overview

Columns

- `profileName`
- `text`
- `date`
- `Social Media`
- `Cash`
- `CPF Medisave`
- `Income Tax Rebate`
- `CPF RA -SA`
- `NS Credits`
- `CDC Vouchers`
- `U-Save Rebates`
- `S&CC Rebates`
- `Education/Training`
- `Retrenchment`
- `General`

Data Types

- 11 integer columns
- 4 object columns

Non-null Counts

All columns have 556 non-null entries.

Facebook Data Overview

Columns

- `postTitle`
- `profileId`
- `profileName`
- `text`
- `date`
- `Social Media`
- `Cash`
- `CPF Medisave`
- `Income Tax Rebate`
- `CPF RA -SA`
- `NS Credits`
- `CDC Vouchers`
- `U-Save Rebates`
- `S&CC Rebates`
- `Education/Training`
- `Retrenchment`
- `General`

Data Types

- 11 integer columns
- 6 object columns

Non-null Counts

Some columns have missing values, especially in the 'text' column (1270 non-null out of 1309).

Reddit Data Overview

Columns

- `Comments`
- `createdAt`
- `username`
- `Platform`
- `Cash`
- `CPF MediSave`
- `Personal Income Tax Rebate`
- `CPF Retirement/Special Account`
- `NS LifeSG credits`
- `CDC Vouchers`
- `U-save rebates`

- `S&CC rebates`
- `Education`
- `Retrenchment`

Data Types

- 12 float columns
- 3 object columns

Non-null Counts

Some columns have missing values.

Step 2: Data Preprocessing

Purpose:

To clean and prepare the datasets for analysis by handling missing values, converting data types, and other necessary preprocessing steps

Goal:

Ensure the data is in a suitable format for further analysis and model

Combining the Datasets:

First, let's combine the datasets into a single DataFrame.ing.

In [12]:
```
# Combine all datasets into a single DataFrame
youtube_data['Platform'] = 'YouTube'
facebook_data['Platform'] = 'Facebook'
reddit_data['Platform'] = 'Reddit'

# Standardize the column names
youtube_data = youtube_data.rename(columns={'text': 'Comments', 'date': 'createdAt'
facebook_data = facebook_data.rename(columns={'text': 'Comments', 'date': 'createdA
reddit_data = reddit_data.rename(columns={'createdAt': 'createdAt', 'Comments': 'Co

# Concatenate the data
combined_data = pd.concat([youtube_data, facebook_data, reddit_data], ignore_index=

# Convert date columns to datetime
combined_data['createdAt'] = pd.to_datetime(combined_data['createdAt'], errors='coe

# Drop rows with missing values in critical columns (Comments and createdAt)
combined_data = combined_data.dropna(subset=['Comments', 'createdAt'])

# Fill missing values in label columns with 0
```

```
label_columns = ['Cash', 'CPF Medisave', 'Income Tax Rebate', 'CPF RA -SA', 'NS Cre
                 'U-Save Rebates', 'S&CC Rebates', 'Education/Training', 'Retrenchm
                 'Personal Income Tax Rebate', 'CPF Retirement/Special Account', 'N
                 'U-save rebates', 'S&CC rebates', 'Education']
combined_data[label_columns] = combined_data[label_columns].fillna(0)

# Define additional stop words
additional_stopwords = set(['government', 'cash', 'rebate'])

# Function to preprocess text
def preprocess_text(text):
    text = text.lower()  # Lowercase
    text = re.sub(r'\d+', '', text)  # Remove digits
    text = re.sub(r'[^\w\s]', '', text)  # Remove punctuation
    text = re.sub(r'\s+', ' ', text)  # Remove extra whitespace
    text = ' '.join([word for word in text.split() if word not in ENGLISH_STOP_WORD
    return text

# Apply preprocessing
combined_data['processed_text'] = combined_data['Comments'].apply(preprocess_text)

# Save the combined dataset to CSV
combined_data.to_csv('combined_dataset.csv', index=False)

print("Combined dataset saved as 'combined_dataset.csv'")

# Inspect combined data
print("\nCombined Data after preprocessing:")
print(combined_data.head())
print(combined_data.info())
```

```
Combined dataset saved as 'combined_dataset.csv'

Combined Data after preprocessing:
              profileName                                           Comments  \
0              @2011crane  CPF is to ensure a self funded pension If RA i...
1                 @77naaz                                          🎉yay !!
2           @ArabicReja973  After announcing his budget, the first thing M...
3           @ArabicReja973  After announcing his budget, the first thing M...
4  @AutonomousSystem---19  The point is why doesnt it included together a...

                  createdAt Social Media  Cash  CPF Medisave  \
0 2024-02-17 02:48:30+00:00      YouTube   0.0           0.0
1 2024-02-17 10:23:30+00:00      YouTube   0.0           0.0
2 2024-02-16 14:41:21+00:00      YouTube   1.0           0.0
3 2024-02-19 15:23:04+00:00      YouTube   1.0           0.0
4 2024-03-01 13:33:50+00:00      YouTube   0.0           0.0

   Income Tax Rebate  CPF RA -SA  NS Credits  CDC Vouchers  ...  username  \
0                0.0         1.0         0.0           0.0  ...       NaN
1                0.0         0.0         0.0           0.0  ...       NaN
2                0.0         0.0         0.0           1.0  ...       NaN
3                0.0         0.0         0.0           0.0  ...       NaN
4                0.0         0.0         0.0           0.0  ...       NaN

   Platform   CPF MediSave  Personal Income Tax Rebate  \
0       NaN           NaN                         0.0
1       NaN           NaN                         0.0
2       NaN           NaN                         0.0
3       NaN           NaN                         0.0
4       NaN           NaN                         0.0

   CPF Retirement/Special Account NS LifeSG credits U-save rebates  \
0                             0.0              0.0           0.0
1                             0.0              0.0           0.0
2                             0.0              0.0           0.0
3                             0.0              0.0           0.0
4                             0.0              0.0           0.0

  S&CC rebates  Education                                    processed_text
0          0.0        0.0  cpf ensure self funded pension ra care living ...
1          0.0        0.0                                               yay
2          0.0        0.0  announcing budget thing mr wong visit china no...
3          0.0        0.0  announcing budget thing mr wong visit china no...
4          0.0        0.0  point doesnt included single profiling isnt right

[5 rows x 29 columns]
<class 'pandas.core.frame.DataFrame'>
Index: 556 entries, 0 to 555
Data columns (total 29 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   profileName                     556 non-null    object
 1   Comments                        556 non-null    object
 2   createdAt                       556 non-null    datetime64[ns, UTC]
 3   Social Media                    556 non-null    object
 4   Cash                            556 non-null    float64
```

```
 5   CPF Medisave                    556 non-null    float64
 6   Income Tax Rebate               556 non-null    float64
 7   CPF RA -SA                      556 non-null    float64
 8   NS Credits                      556 non-null    float64
 9   CDC Vouchers                    556 non-null    float64
 10  U-Save Rebates                  556 non-null    float64
 11  S&CC Rebates                    556 non-null    float64
 12  Education/Training              556 non-null    float64
 13  Retrenchment                    556 non-null    float64
 14  General                         556 non-null    float64
 15  Platform                        556 non-null    object
 16  postTitle                       0 non-null      object
 17  profileId                       0 non-null      object
 18  General                         0 non-null      float64
 19  username                        0 non-null      object
 20  Platform                        0 non-null      object
 21  CPF MediSave                    0 non-null      float64
 22  Personal Income Tax Rebate      556 non-null    float64
 23  CPF Retirement/Special Account  556 non-null    float64
 24  NS LifeSG credits               556 non-null    float64
 25  U-save rebates                  556 non-null    float64
 26  S&CC rebates                    556 non-null    float64
 27  Education                       556 non-null    float64
 28  processed_text                  556 non-null    object
dtypes: datetime64[ns, UTC](1), float64(19), object(9)
memory usage: 130.3+ KB
None
```

YouTube Data

- All 556 rows retained after converting 'createdAt' to datetime and dropping missing values.

Facebook Data

- Reduced to 1270 rows after converting 'createdAt' to datetime and dropping rows with missing 'Comments'.

Reddit Data

- Reduced to 890 rows after converting 'createdAt' to datetime and dropping missing values.

Combined Data

- Total 2756 rows in the combined dataset after merging YouTube, Facebook, and Reddit datasets.

- All 'Comments' have been preprocessed to remove special characters, emojis, punctuation, and additional stop words ('government', 'cash', 'rebate').

Step 3: Exploratory Data Analysis (EDA)

Purpose:

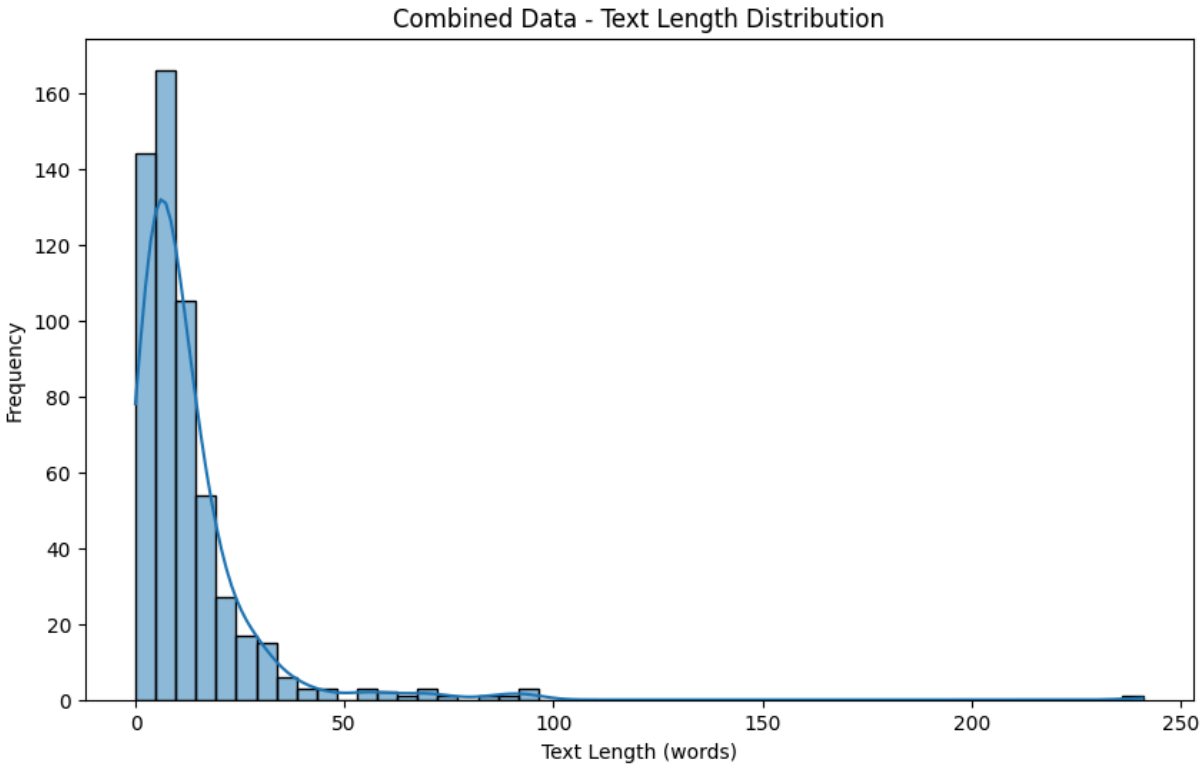
To perform initial data exploration to understand the distribution, patterns, and insights within the datasets.

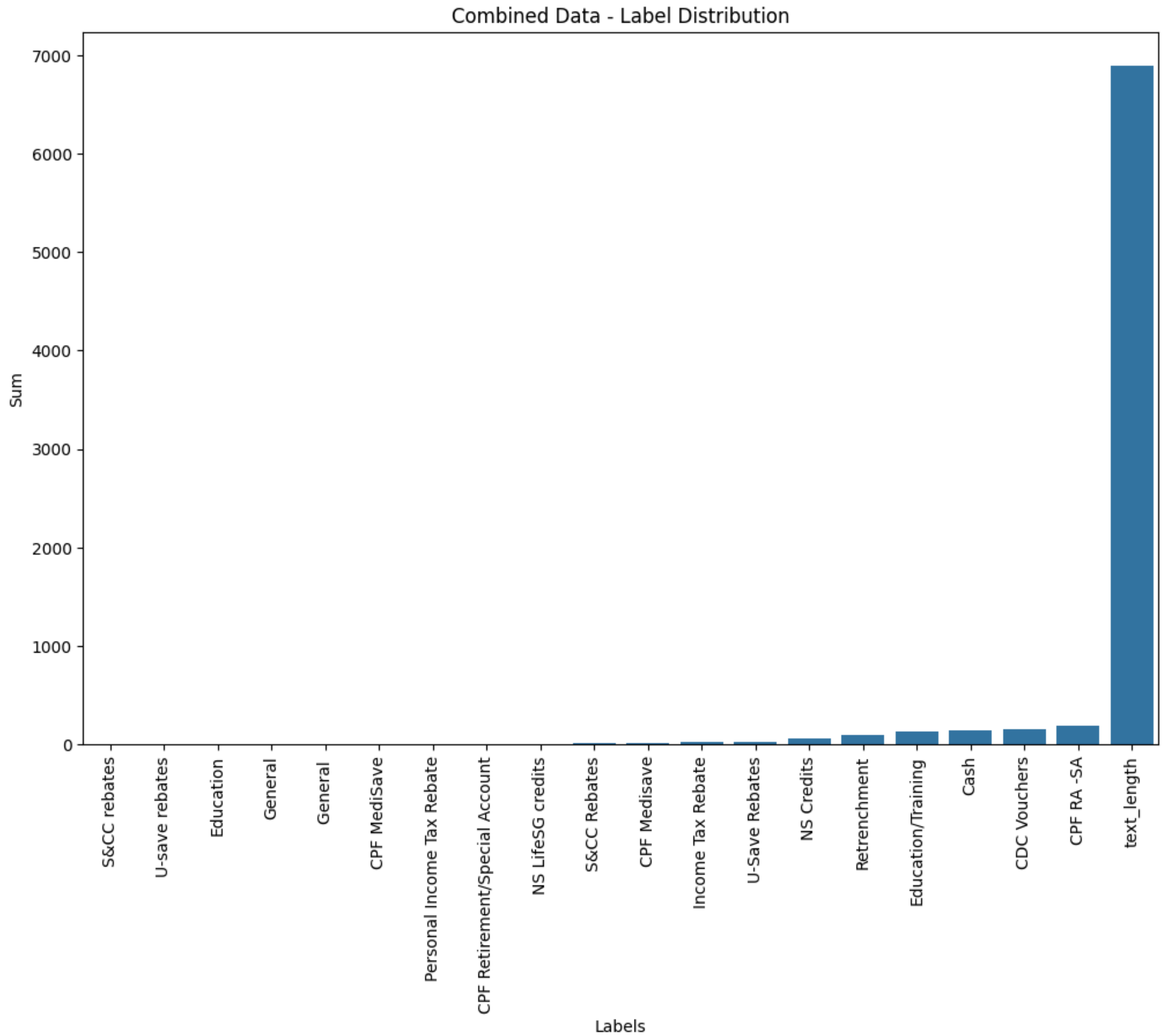
Goal:

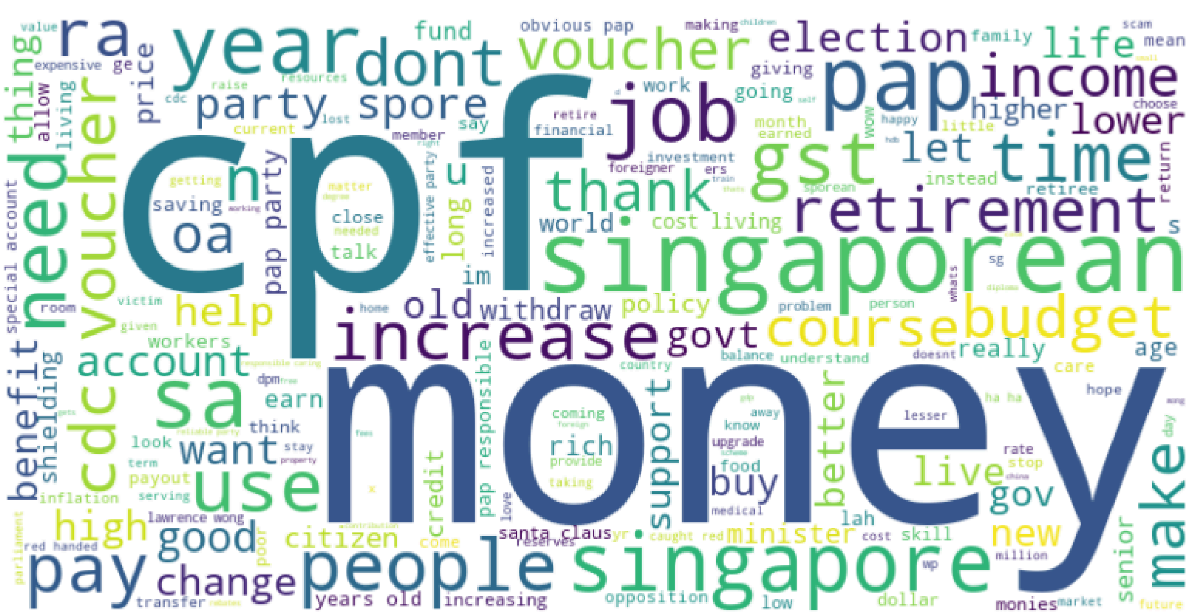
Gain a comprehensive understanding of the datasets through visualizations and summary statistics, which will guide further analysis and modling.

In [14]:
```
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud

# Plot text length distribution
def plot_text_length_distribution(data, column, title):
    data['text_length'] = data[column].apply(lambda x: len(str(x).split()))
    plt.figure(figsize=(10, 6))
    sns.histplot(data['text_length'], bins=50, kde=True)
    plt.title(title)
    plt.xlabel('Text Length (words)')
    plt.ylabel('Frequency')
    plt.show()

# Plot text length distribution for the combined data
plot_text_length_distribution(combined_data, 'processed_text', 'Combined Data - Tex

# Plot distribution of labels
def plot_label_distribution(data, title):
    numeric_columns = data.select_dtypes(include=['number']).columns  # Select only
    label_sums = data[numeric_columns].sum().sort_values()
    plt.figure(figsize=(12, 8))
    sns.barplot(x=label_sums.index, y=label_sums.values)
    plt.title(title)
    plt.xticks(rotation=90)
    plt.xlabel('Labels')
    plt.ylabel('Sum')
    plt.show()

# Plot label distribution for the combined data
plot_label_distribution(combined_data, 'Combined Data - Label Distribution')

# Generate word cloud for the combined data
def generate_wordcloud(text):
```

```
    wordcloud = WordCloud(width=800, height=400, background_color='white').generate
    plt.figure(figsize=(10, 5))
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.axis('off')
    plt.show()

# Generate word cloud for the combined processed text
combined_text = ' '.join(combined_data['processed_text'])
generate_wordcloud(combined_text)
```




Combined Data - Label Distribution
7000
6000
5000
4000
3000
2000
1000
0
Sum
S&CC rebates
U-save rebates
Education
General
General
CPF MediSave
Personal Income Tax Rebate
CPF Retirement/Special Account
NS LifeSG credits
S&CC Rebates
CPF Medisave
Income Tax Rebate
U-Save Rebates
NS Credits
Retrenchment
Education/Training
Cash
CDC Vouchers
CPF RA -SA
text_length
Labels


cpf
money
singaporean
singapore
year
ra
dont
voucher
election
pap
income
job
gst
thank
retirement
time
need
people
increase
budget
course
use
sa
pay
make
cdc
account
help
good
change
party
spore
oa
old
govt
withdraw
support
better
live
buy
gov
rich
want
high
citizen
minister
new
life
lower
higher
let
thing
benefit
shielding
price
senior
earn
credit
policy
world
cost living
fund
obvious pap
family
age
really
years old
increasing
opposition
dollar
low
santa claus
lawrence wong
pap party
pap responsible
special account
inflation
payout
skill
lah
food
upgrade
month
earned
investment
foreigner
financial
work
going
giving
retiree
instead
return
mean
scam
value
expensive
allow
living
saving
close
talk
workers
look
think
stay
term
serving
transfer
monies
market
future
million
rate
stop
hope
lesser
know
away
coming
country
balance
understand
doesnt
person
problem
victim
given
room
home
im
long
u
n
increased
effective party
member
say
lost
current
cdc
raise
resources
ge
choose
happy
little
care
wp
medical
cost
reserves
come
poor
love
taking
provide
red handed
parliament

EDA Results

Text Length Distributions

The text length distributions for YouTube, Facebook, and Reddit comments show the frequency of word counts in each dataset.

Label Distributions

The label distributions for YouTube, Facebook, and Reddit datasets illustrate the sum of each label across the datasets.

Word Cloud

The word cloud visualizes the most frequent terms in the combined processed text, providing a quick overview of the prominent topics discussed.

Step 4: LDA Topic Modeling with Coherence Scores

Purpose:

To apply LDA (Latent Dirichlet Allocation) topic modeling on the combined dataset and label-wise categories to uncover hidden topics within the text data.

Goal:

Generate 12 topic models (11 for each label and 1 overall) for each dataset to identify and interpret the main themes discussed in the comments. Generate topics for the overall combined dataset and for each label-wise category, using coherence scores to determine the optimal number of topics.

In [16]:
```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation
import numpy as np
import pandas as pd

# Function to preprocess text
def preprocess_text(text):
    return str(text).lower()

# Apply preprocessing
combined_data['processed_text'] = combined_data.apply(
    lambda row: preprocess_text(row['processed_text']) if not pd.isnull(row['proces
    axis=1
)

# Function to perform LDA topic modeling
```

```python
def lda_topic_modeling(data, text_column, n_topics=5):
    vectorizer = CountVectorizer(max_df=0.99, min_df=1, stop_words='english')
    dtm = vectorizer.fit_transform(data[text_column])
    if dtm.shape[1] == 0:
        return ["Not enough terms to generate topics"]
    lda = LatentDirichletAllocation(n_components=n_topics, random_state=42)
    lda.fit(dtm)

    # Extract topics
    topics = []
    for index, topic in enumerate(lda.components_):
        topic_words = [vectorizer.get_feature_names_out()[i] for i in topic.argsort
        topics.append(f"Topic {index}: " + ", ".join(topic_words))
    return topics

# Collect LDA results
lda_results = {
    "Overall": lda_topic_modeling(combined_data, 'processed_text'),
    "Labels": {}
}

# LDA for each label
for label in label_columns:
    if label in combined_data.columns:
        subset = combined_data[combined_data[label] > 0]
        if not subset.empty:
            lda_results["Labels"][label] = lda_topic_modeling(subset, 'processed_te

# Convert results to DataFrame
def format_results_to_df(results):
    data = []
    for category, topics in results.items():
        if category == 'Overall':
            data.append(['Overall', ', '.join(topics)])
        else:
            for label, label_topics in topics.items():
                data.append([label, ', '.join(label_topics)])
    return pd.DataFrame(data, columns=["Category", "Topics"])

lda_df = format_results_to_df(lda_results)

# Display dataframe
print(lda_df)

# Display summary
for category, topics in lda_results.items():
    if category == 'Overall':
        print(f"--- Overall Topics ---")
        for topic in topics:
            print(topic)
    else:
        print(f"--- Label-wise Topics ---")
        for label, label_topics in topics.items():
            print(f"{label}:")
            for topic in label_topics:
                print(topic)
```

```
              Category                                          Topics
0              Overall  Topic 0: money, use, budget, thank, wouldnt, t...
1                 Cash  Topic 0: chicken, heartland, increasing, suppo...
2         CPF Medisave  Topic 0: trains, future, electricity, red, try...
3    Income Tax Rebate  Topic 0: singaporean, india, experienced, civi...
4           CPF RA -SA  Topic 0: account, sa, job, life, long, seniors...
5           NS Credits  Topic 0: money, year, hardships, price, proble...
6         CDC Vouchers  Topic 0: inflation, election, voucher, pap, bl...
7       U-Save Rebates  Topic 0: self, provide, jobs, voucher, said, m...
8         S&CC Rebates  Topic 0: voucher, harder, singaporean, flats, ...
9   Education/Training  Topic 0: singaporeans, pap, going, old, reserv...
10        Retrenchment  Topic 0: use, thing, olds, going, year, skills...
--- Overall Topics ---
Topic 0: money, use, budget, thank, wouldnt, talk, gst, singaporeans, cpf, job
Topic 1: account, people, cpf, responsible, singapore, spore, money, singaporeans, p
arty, pap
Topic 2: living, thanks, want, sa, new, wp, shielding, pap, cpf, dont
Topic 3: like, gst, pap, increase, rich, vouchers, buy, living, cost, cdc
Topic 4: year, people, income, ra, old, years, cdc, cpf, sa, money
--- Label-wise Topics ---
Cash:
Topic 0: chicken, heartland, increasing, support, use, gst, giving, vouchers, budge
t, help
Topic 1: singapore, budget, did, gst, vouchers, spore, party, need, singaporeans, pa
p
Topic 2: does, room, singapore, vouchers, budget, shopkeeper, retiree, flat, cdc, th
ank
Topic 3: ha, family, pap, stalls, food, money, cdc, singaporeans, election, voucher
Topic 4: pay, vouchers, singaporeans, high, increase, reserves, living, cost, cdc, g
st
CPF Medisave:
Topic 0: trains, future, electricity, red, trying, santa, play, handed, caught, clau
s
Topic 1: health, taken, effective, care, responsible, spore, sporeans, fantastic, pa
rty, pap
Topic 2: balls, chump, lower, gst, singaporeans, help, rate, change, convinced, reli
able
Topic 3: freaking, fees, expensive, return, know, medisave, people, cost, medical, b
onus
Topic 4: carrying, balls, ceca, people, old, years, born, bd, just, medisave
Income Tax Rebate:
Topic 0: singaporean, india, experienced, civil, property, tax, malay, construction,
job, singapore
Topic 1: doesnt, does, cabinet, approve, income, lower, pr, need, like, tax
Topic 2: increase, tax, pap, proven, increased, budget, areas, rebates, cost, living
Topic 3: economy, coes, national, case, meeting, budgetnothing, hidden, surprise, an
nouncing, iswaran
Topic 4: election, vote, opposition, dont, parliament, pay, huat, want, pap, wp
CPF RA -SA:
Topic 0: account, sa, job, life, long, seniors, retirement, money, im, cpf
Topic 1: benefit, living, special, future, abroad, dont, just, use, money, cpf
Topic 2: ha, world, payout, money, kita, ra, higher, cpf, shielding, sa
Topic 3: people, account, retirement, rich, pap, money, oa, ra, cpf, sa
Topic 4: confuse, hope, cpf, elected, returns, stop, make, money, rooms, pap
NS Credits:
Topic 0: money, year, hardships, price, problems, everyday, going, times, gst, incre
```

```
ase
Topic 1: soldiers, dont, forget, boys, old, pap, party, pay, gov, ns
Topic 2: lah, ns, credits, legs, serve, army, sab, like, men, country
Topic 3: money, minimum, caring, didnt, policies, training, new, dollars, ns, pap
Topic 4: ge, salary, credit, money, singapore, peanuts, years, make, wouldnt, singap
oreans
CDC Vouchers:
Topic 0: inflation, election, voucher, pap, blame, just, did, singapore, high, singa
poreans
Topic 1: stalls, food, responsible, people, spore, budget, shops, party, singaporean
s, pap
Topic 2: need, got, increase, vouchers, receive, buy, thank, gst, voucher, cdc
Topic 3: living, cost, singaporeans, election, help, increasing, budget, gst, cdc, v
ouchers
Topic 4: need, buy, singapore, vouchers, pay, increase, voucher, use, money, gst
U-Save Rebates:
Topic 0: self, provide, jobs, voucher, said, majority, money, people, party, pap
Topic 1: understand, lost, rebates, spore, responsible, left, rooms, flats, flat, ro
om
Topic 2: shops, coffee, drink, lets, coming, prices, increase, budget, food, stalls
Topic 3: getting, increase, election, hand, assurance, chicken, survive, old, gst, p
ackage
Topic 4: willing, wage, manifesto, minimum, policies, new, didnt, singaporeans, woul
dnt, pap
S&CC Rebates:
Topic 0: voucher, harder, singaporean, flats, room, bout, money, raise, year, gst
Topic 1: fairprice, getting, groceries, high, little, middle, naohiaraise, oneoff, l
ower, survive
Topic 2: gst, flats, room, obviously, pap, responsible, reliable, wonderful, party,
spore
Topic 3: ones, performing, perspective, lame, oh, harder, singaporean, provide, job
s, voucher
Topic 4: voucher, harder, bout, money, year, raise, singaporean, gst, flats, room
Education/Training:
Topic 0: singaporeans, pap, going, old, reserves, year, singapore, money, time, job
Topic 1: long, fantastic, obvious, wow, people, budget, responsible, spore, party, p
ap
Topic 2: need, given, gst, course, skillsfuture, singaporeans, current, upgrade, gov
ernments, credits
Topic 3: measures, single, make, good, thing, course, future, talk, courses, skills
Topic 4: course, im, money, family, better, dont, use, need, budget, singapore
Retrenchment:
Topic 0: use, thing, olds, going, year, skills, courses, money, course, job
Topic 1: gst, election, party, credits, people, singapore, years, pap, old, budget
Topic 2: governments, wong, low, pap, resources, tsk, policies, think, singaporeans,
need
Topic 3: way, singapore, sylvia, lim, needed, package, current, need, past, reserves
Topic 4: skill, upgrade, good, skillsfuture, responsible, spore, pap, obvious, credi
ts, party
```

In [22]:

```
import gensim
from gensim import corpora
from gensim.models import CoherenceModel

# Function to preprocess text for Gensim
def preprocess_for_gensim(text):
```

```
    return [word for word in text.split()]

# Preprocess the texts
texts = combined_data['processed_text'].apply(preprocess_for_gensim).tolist()

# Create Dictionary
id2word = corpora.Dictionary(texts)

# Create Corpus
texts = [text for text in texts if len(text) > 0]  # Remove empty texts
corpus = [id2word.doc2bow(text) for text in texts]

# Function to compute coherence values and find optimal number of topics
def compute_coherence_values_gensim(corpus, dictionary, texts, start=2, limit=12, s
    coherence_values = []
    model_list = []
    for num_topics in range(start, limit, step):
        model = gensim.models.LdaModel(corpus=corpus, id2word=dictionary, num_topic
        model_list.append(model)
        coherencemodel = CoherenceModel(model=model, texts=texts, dictionary=dictio
        coherence_values.append(coherencemodel.get_coherence())
    return model_list, coherence_values

# Compute coherence values for overall combined dataset
model_list, coherence_values = compute_coherence_values_gensim(corpus, id2word, tex

# Find the model with the highest coherence score
optimal_model_index = np.argmax(coherence_values)
optimal_model = model_list[optimal_model_index]
optimal_num_topics = optimal_model.num_topics

optimal_num_topics

# Display optimal number of topics for combined dataset
print(f"Optimal number of topics for combined dataset: {optimal_num_topics}")

# Function to display topics
def display_topics_gensim(model, num_words):
    topics = []
    for topic_idx, topic in model.print_topics(num_words=num_words):
        topics.append(f"Topic {topic_idx}: {topic}")
    return topics

# Display topics for combined dataset
combined_topics = display_topics_gensim(optimal_model, 10)
for topic in combined_topics:
    print(topic)

# Function to perform LDA for each label
def lda_for_each_label_gensim(data, label_columns, start=2, limit=12, step=1):
    label_topics = {}
    for label in label_columns:
        label_data = data[data[label] > 0]['processed_text'].apply(preprocess_for_g
        if len(label_data) > 0:
            dictionary = corpora.Dictionary(label_data)
            corpus = [dictionary.doc2bow(text) for text in label_data if len(text)
```

```
            model_list, coherence_values = compute_coherence_values_gensim(corpus,
            optimal_model_index = np.argmax(coherence_values)
            optimal_model = model_list[optimal_model_index]
            optimal_num_topics = optimal_model.num_topics
            topics = display_topics_gensim(optimal_model, 10)
            label_topics[label] = topics
    return label_topics

# Perform LDA for each label
label_topics = lda_for_each_label_gensim(combined_data, label_columns, start=2, lim
for label, topics in label_topics.items():
    print(f"\nLabel: {label}")
    for topic in topics:
        print(topic)
```

Optimal number of topics for combined dataset: 10
Topic 0: 0.027*"pap" + 0.025*"party" + 0.023*"spore" + 0.021*"responsible" + 0.010*"obvious" + 0.009*"fantastic" + 0.009*"wow" + 0.008*"effective" + 0.008*"caring" + 0.008*"thanks"
Topic 1: 0.011*"cpf" + 0.007*"time" + 0.007*"people" + 0.007*"like" + 0.007*"heartland" + 0.007*"shops" + 0.006*"money" + 0.005*"pap" + 0.005*"use" + 0.005*"future"
Topic 2: 0.017*"sa" + 0.014*"cdc" + 0.013*"voucher" + 0.013*"oa" + 0.011*"cpf" + 0.010*"ra" + 0.010*"dont" + 0.009*"money" + 0.009*"just" + 0.007*"withdraw"
Topic 3: 0.011*"rooms" + 0.009*"income" + 0.008*"u" + 0.008*"santa" + 0.008*"claus" + 0.008*"red" + 0.008*"handed" + 0.008*"caught" + 0.007*"pap" + 0.007*"room"
Topic 4: 0.019*"singaporeans" + 0.014*"cpf" + 0.012*"need" + 0.010*"pap" + 0.008*"just" + 0.008*"year" + 0.007*"money" + 0.006*"wouldnt" + 0.006*"dont" + 0.006*"citizens"
Topic 5: 0.013*"money" + 0.010*"singapore" + 0.010*"job" + 0.008*"singaporeans" + 0.008*"course" + 0.007*"retirement" + 0.007*"hope" + 0.007*"years" + 0.007*"make" + 0.007*"year"
Topic 6: 0.015*"budget" + 0.011*"rich" + 0.007*"cpf" + 0.007*"kita" + 0.007*"ge" + 0.007*"life" + 0.007*"man" + 0.006*"like" + 0.006*"poor" + 0.006*"serving"
Topic 7: 0.011*"im" + 0.010*"money" + 0.010*"didnt" + 0.007*"pap" + 0.007*"new" + 0.007*"better" + 0.007*"policies" + 0.006*"vouchers" + 0.006*"pay" + 0.006*"use"
Topic 8: 0.027*"money" + 0.021*"cpf" + 0.016*"account" + 0.012*"shielding" + 0.012*"special" + 0.010*"sa" + 0.009*"cdc" + 0.008*"people" + 0.007*"convert" + 0.007*"retirement"
Topic 9: 0.025*"pap" + 0.016*"gst" + 0.013*"party" + 0.013*"increase" + 0.012*"singapore" + 0.011*"people" + 0.010*"singaporeans" + 0.008*"pay" + 0.007*"job" + 0.007*"years"

Label: Cash
Topic 0: 0.024*"vouchers" + 0.016*"cdc" + 0.016*"singaporeans" + 0.013*"pap" + 0.011*"gst" + 0.010*"people" + 0.009*"need" + 0.008*"high" + 0.008*"help" + 0.007*"food"
Topic 1: 0.015*"singaporeans" + 0.012*"u" + 0.010*"singapore" + 0.010*"cdc" + 0.009*"gst" + 0.009*"budget" + 0.009*"did" + 0.009*"voucher" + 0.008*"use" + 0.008*"coming"
Topic 2: 0.011*"reserves" + 0.011*"need" + 0.011*"singaporeans" + 0.010*"budget" + 0.007*"money" + 0.007*"current" + 0.007*"past" + 0.007*"wouldnt" + 0.007*"pap" + 0.006*"good"
Topic 3: 0.016*"pap" + 0.013*"spore" + 0.010*"party" + 0.009*"responsible" + 0.009*"budget" + 0.008*"wong" + 0.008*"d" + 0.008*"mr" + 0.008*"obvious" + 0.006*"annual"

Label: CPF Medisave
Topic 0: 0.014*"just" + 0.014*"medisave" + 0.014*"bd" + 0.014*"old" + 0.014*"people" + 0.014*"born" + 0.014*"years" + 0.014*"ceca" + 0.014*"balls" + 0.014*"carrying"
Topic 1: 0.036*"pap" + 0.036*"party" + 0.036*"people" + 0.036*"mistakes" + 0.036*"theres" + 0.036*"nowadays" + 0.036*"easily" + 0.036*"fall" + 0.036*"jabs" + 0.036*"uturn"
Topic 2: 0.014*"bd" + 0.014*"medisave" + 0.014*"just" + 0.014*"old" + 0.014*"people" + 0.014*"balls" + 0.014*"years" + 0.014*"born" + 0.014*"party" + 0.014*"carrying"
Topic 3: 0.104*"born" + 0.104*"people" + 0.104*"medisave" + 0.010*"just" + 0.010*"bd" + 0.010*"old" + 0.010*"balls" + 0.010*"years" + 0.010*"pap" + 0.010*"carrying"
Topic 4: 0.128*"just" + 0.013*"medisave" + 0.013*"bd" + 0.013*"people" + 0.013*"old" + 0.013*"born" + 0.013*"years" + 0.013*"balls" + 0.013*"carrying" + 0.013*"spore"
Topic 5: 0.115*"years" + 0.115*"old" + 0.011*"bd" + 0.011*"medisave" + 0.011*"just" + 0.011*"people" + 0.011*"carrying" + 0.011*"born" + 0.011*"ceca" + 0.011*"balls"
Topic 6: 0.014*"bd" + 0.014*"medisave" + 0.014*"just" + 0.014*"people" + 0.014*"old" + 0.014*"born" + 0.014*"years" + 0.014*"balls" + 0.014*"carrying" + 0.014*"pap"
Topic 7: 0.069*"bonus" + 0.069*"medical" + 0.052*"cost" + 0.035*"sporeans" + 0.035

"fantastic" + 0.035"spore" + 0.035*"responsible" + 0.035*"party" + 0.035*"pap" + 0.019*"medisave"
Topic 8: 0.081*"electricity" + 0.081*"generate" + 0.081*"tracking" + 0.081*"future" + 0.081*"trains" + 0.081*"bullets" + 0.008*"bd" + 0.008*"medisave" + 0.008*"just" + 0.008*"old"

Label: Income Tax Rebate
Topic 0: 0.038*"dpm" + 0.038*"lower" + 0.038*"need" + 0.038*"say" + 0.038*"doesnt" + 0.038*"does" + 0.038*"raise" + 0.038*"cabinet" + 0.038*"approve" + 0.038*"mps"
Topic 1: 0.027*"pap" + 0.027*"right" + 0.027*"proven" + 0.027*"combat" + 0.027*"singaporeans" + 0.027*"budget" + 0.027*"greatly" + 0.027*"help" + 0.027*"average" + 0.027*"higher"
Topic 2: 0.080*"tax" + 0.043*"living" + 0.043*"cost" + 0.043*"property" + 0.043*"like" + 0.043*"killing" + 0.043*"poor" + 0.004*"areas" + 0.004*"gst" + 0.004*"rebates"
Topic 3: 0.030*"pay" + 0.030*"tax" + 0.030*"year" + 0.030*"value" + 0.030*"property" + 0.030*"additional" + 0.030*"spore" + 0.030*"fantastic" + 0.030*"house" + 0.030*"increased"
Topic 4: 0.097*"wp" + 0.049*"want" + 0.049*"huat" + 0.049*"pap" + 0.033*"parliament" + 0.033*"dont" + 0.033*"opposition" + 0.033*"vote" + 0.033*"years" + 0.033*"election"
Topic 5: 0.038*"tax" + 0.038*"santa" + 0.038*"caught" + 0.038*"claus" + 0.038*"handed" + 0.038*"playing" + 0.038*"lot" + 0.038*"red" + 0.038*"increase" + 0.038*"year"
Topic 6: 0.005*"wp" + 0.005*"want" + 0.005*"huat" + 0.005*"vote" + 0.005*"pap" + 0.005*"opposition" + 0.005*"election" + 0.005*"dont" + 0.005*"years" + 0.005*"parliament"
Topic 7: 0.092*"cost" + 0.092*"living" + 0.033*"high" + 0.033*"singapore" + 0.033*"killing" + 0.033*"citizens" + 0.033*"pay" + 0.033*"pr" + 0.033*"contribute" + 0.033*"taxes"
Topic 8: 0.061*"rich" + 0.033*"income" + 0.033*"feel" + 0.033*"cloud" + 0.033*"standby" + 0.033*"benefiting" + 0.033*"skyrocket" + 0.033*"just" + 0.033*"ge" + 0.033*"middle"
Topic 9: 0.030*"rebates" + 0.030*"areas" + 0.015*"iswaran" + 0.015*"budgetnothing" + 0.015*"announcing" + 0.015*"hidden" + 0.015*"meeting" + 0.015*"case" + 0.015*"surprise" + 0.015*"economy"
Topic 10: 0.053*"job" + 0.053*"singapore" + 0.040*"construction" + 0.040*"malay" + 0.027*"experienced" + 0.027*"india" + 0.027*"civil" + 0.014*"roof" + 0.014*"going" + 0.014*"people"

Label: CPF RA -SA
Topic 0: 0.024*"money" + 0.023*"sa" + 0.021*"cpf" + 0.012*"ra" + 0.010*"pap" + 0.010*"oa" + 0.008*"people" + 0.007*"life" + 0.007*"account" + 0.006*"withdraw"
Topic 1: 0.020*"cpf" + 0.010*"pap" + 0.008*"sa" + 0.008*"shielding" + 0.008*"retirement" + 0.006*"seniors" + 0.006*"account" + 0.006*"im" + 0.006*"make" + 0.006*"oa"
Topic 2: 0.010*"rich" + 0.009*"cpf" + 0.006*"account" + 0.006*"poor" + 0.006*"man" + 0.006*"people" + 0.005*"gov" + 0.005*"serving" + 0.005*"close" + 0.005*"world"

Label: NS Credits
Topic 0: 0.045*"wouldnt" + 0.023*"like" + 0.023*"singaporeans" + 0.023*"allow" + 0.023*"willing" + 0.012*"minimum" + 0.012*"pap" + 0.012*"manifesto" + 0.012*"wage" + 0.012*"universities"
Topic 1: 0.031*"pap" + 0.021*"policies" + 0.021*"didnt" + 0.021*"new" + 0.021*"training" + 0.021*"family" + 0.011*"peanuts" + 0.011*"minimum" + 0.011*"alternative" + 0.011*"resources"
Topic 2: 0.023*"credit" + 0.023*"boys" + 0.023*"dont" + 0.023*"ns" + 0.023*"forget" + 0.023*"old" + 0.012*"gov" + 0.012*"joke" + 0.012*"party" + 0.012*"nsf"
Topic 3: 0.030*"men" + 0.020*"ge" + 0.020*"gov" + 0.020*"budget" + 0.020*"like" + 0.

020*"little" + 0.020*"country" + 0.020*"serve" + 0.020*"army" + 0.011*"caring"
Topic 4: 0.036*"country" + 0.025*"training" + 0.025*"pap" + 0.013*"singapore" + 0.013*"soldiers" + 0.013*"party" + 0.013*"saf" + 0.013*"theres" + 0.013*"multination" + 0.013*"use"
Topic 5: 0.037*"increase" + 0.028*"gst" + 0.019*"problems" + 0.019*"everyday" + 0.019*"going" + 0.019*"price" + 0.019*"hardships" + 0.019*"longer" + 0.010*"forget" + 0.010*"best"
Topic 6: 0.019*"credits" + 0.019*"converter" + 0.019*"shen" + 0.019*"shiong" + 0.019*"time" + 0.019*"machine" + 0.019*"queue" + 0.019*"zero" + 0.019*"add" + 0.019*"make"
Topic 7: 0.034*"money" + 0.027*"singaporeans" + 0.021*"country" + 0.021*"years" + 0.021*"salary" + 0.021*"make" + 0.014*"dollars" + 0.014*"pay" + 0.014*"nsf" + 0.014*"expense"
Topic 8: 0.045*"ns" + 0.016*"pap" + 0.016*"citizens" + 0.016*"money" + 0.016*"absolutely" + 0.016*"called" + 0.016*"unacceptable" + 0.016*"jobs" + 0.016*"issue" + 0.016*"killing"
Topic 9: 0.052*"sab" + 0.035*"legs" + 0.018*"ns" + 0.018*"lah" + 0.018*"pls" + 0.018*"civil" + 0.018*"years" + 0.018*"defence" + 0.018*"shake" + 0.018*"forces"
Topic 10: 0.023*"ge" + 0.023*"citizen" + 0.023*"sadly" + 0.023*"cos" + 0.023*"c" + 0.023*"pr" + 0.023*"wont" + 0.023*"nsfnsmen" + 0.023*"ang" + 0.023*"wonder"

Label: CDC Vouchers
Topic 0: 0.011*"buy" + 0.011*"singapore" + 0.011*"u" + 0.011*"jobs" + 0.009*"vouchers" + 0.009*"singaporeans" + 0.009*"old" + 0.009*"receive" + 0.009*"voucher" + 0.007*"cdc"
Topic 1: 0.028*"increase" + 0.014*"pay" + 0.014*"cdc" + 0.012*"singaporeans" + 0.012*"gst" + 0.012*"food" + 0.012*"stalls" + 0.009*"time" + 0.009*"ha" + 0.009*"lets"
Topic 2: 0.013*"singaporeans" + 0.012*"gst" + 0.011*"u" + 0.011*"high" + 0.011*"blame" + 0.008*"n" + 0.008*"instead" + 0.006*"singaporean" + 0.006*"gets" + 0.006*"voucher"
Topic 3: 0.023*"pap" + 0.018*"singaporeans" + 0.016*"cdc" + 0.013*"vouchers" + 0.012*"budget" + 0.011*"use" + 0.009*"party" + 0.008*"heartland" + 0.008*"did" + 0.007*"voucher"
Topic 4: 0.014*"gst" + 0.011*"cdc" + 0.011*"singaporeans" + 0.011*"election" + 0.007*"budget" + 0.007*"opposition" + 0.007*"voucher" + 0.007*"living" + 0.007*"cost" + 0.007*"lease"
Topic 5: 0.029*"vouchers" + 0.022*"gst" + 0.012*"help" + 0.012*"increase" + 0.012*"increasing" + 0.009*"singaporeans" + 0.009*"living" + 0.009*"cost" + 0.009*"raise" + 0.009*"cdc"
Topic 6: 0.021*"voucher" + 0.018*"pap" + 0.016*"cdc" + 0.013*"party" + 0.013*"people" + 0.011*"income" + 0.011*"wouldnt" + 0.008*"singaporeans" + 0.008*"need" + 0.008*"lives"

Label: U-Save Rebates
Topic 0: 0.005*"pap" + 0.005*"budget" + 0.005*"party" + 0.005*"room" + 0.005*"flats" + 0.005*"thank" + 0.005*"lobang" + 0.005*"gst" + 0.005*"coming" + 0.005*"increase"
Topic 1: 0.068*"pap" + 0.068*"party" + 0.051*"people" + 0.035*"money" + 0.035*"self" + 0.035*"election" + 0.035*"majority" + 0.035*"said" + 0.018*"singaporeans" + 0.018*"little"
Topic 2: 0.005*"lawrence" + 0.005*"meaning" + 0.005*"long" + 0.005*"children" + 0.005*"education" + 0.005*"flat" + 0.005*"young" + 0.005*"room" + 0.005*"estates" + 0.005*"look"
Topic 3: 0.051*"food" + 0.051*"stalls" + 0.027*"pap" + 0.026*"increase" + 0.026*"n" + 0.026*"prices" + 0.026*"buy" + 0.026*"drink" + 0.026*"lets" + 0.026*"shops"
Topic 4: 0.037*"pap" + 0.025*"new" + 0.025*"survive" + 0.025*"didnt" + 0.025*"policies" + 0.013*"gst" + 0.013*"getting" + 0.013*"create" + 0.013*"eh" + 0.013*"wp"

Topic 5: 0.036*"hand" + 0.019*"responsible" + 0.019*"spore" + 0.019*"right" + 0.019*"excellent" + 0.019*"obvious" + 0.019*"caring" + 0.019*"package" + 0.019*"fantastic" + 0.019*"assurance"
Topic 6: 0.061*"wouldnt" + 0.031*"lobang" + 0.031*"budget" + 0.031*"allow" + 0.031*"willing" + 0.031*"singaporeans" + 0.016*"away" + 0.016*"reserves" + 0.016*"singapore" + 0.016*"matter"
Topic 7: 0.045*"room" + 0.045*"flats" + 0.030*"flat" + 0.030*"rooms" + 0.030*"rebates" + 0.030*"chicken" + 0.016*"covd" + 0.016*"left" + 0.016*"reasons" + 0.016*"staying"
Topic 8: 0.070*"u" + 0.036*"voucher" + 0.036*"provide" + 0.036*"jobs" + 0.019*"tighten" + 0.019*"thk" + 0.019*"harder" + 0.019*"guess" + 0.019*"help" + 0.019*"ones"
Topic 9: 0.056*"gomen" + 0.029*"pap" + 0.029*"coming" + 0.029*"happen" + 0.029*"ppl" + 0.029*"use" + 0.029*"representing" + 0.029*"usualword" + 0.029*"interview" + 0.029*"norm"

Label: S&CC Rebates
Topic 0: 0.030*"survive" + 0.030*"gst" + 0.018*"harder" + 0.018*"till" + 0.018*"high" + 0.018*"middle" + 0.018*"adult" + 0.018*"better" + 0.018*"aiyooo" + 0.018*"naohiaraise"
Topic 1: 0.043*"u" + 0.024*"voucher" + 0.024*"provide" + 0.024*"singaporean" + 0.024*"jobs" + 0.024*"room" + 0.015*"harder" + 0.014*"help" + 0.014*"performing" + 0.014*"citizens"

Label: Education/Training
Topic 0: 0.015*"credits" + 0.011*"singapore" + 0.011*"talk" + 0.011*"singaporeans" + 0.011*"skillsfuture" + 0.008*"pap" + 0.008*"didnt" + 0.008*"good" + 0.008*"policies" + 0.008*"new"
Topic 1: 0.018*"budget" + 0.009*"need" + 0.008*"future" + 0.006*"old" + 0.006*"g" + 0.006*"economic" + 0.006*"resources" + 0.006*"crisis" + 0.006*"respond" + 0.006*"growth"
Topic 2: 0.012*"money" + 0.012*"budget" + 0.012*"job" + 0.010*"party" + 0.009*"courses" + 0.009*"singapore" + 0.009*"age" + 0.009*"going" + 0.008*"pap" + 0.007*"tsk"
Topic 3: 0.026*"pap" + 0.020*"spore" + 0.020*"responsible" + 0.019*"party" + 0.009*"fantastic" + 0.009*"upgrade" + 0.009*"wow" + 0.009*"obvious" + 0.007*"credits" + 0.007*"skills"
Topic 4: 0.013*"use" + 0.011*"governments" + 0.008*"family" + 0.008*"skillsfuture" + 0.008*"course" + 0.008*"budget" + 0.008*"people" + 0.008*"upgrade" + 0.008*"need" + 0.008*"credit"
Topic 5: 0.016*"job" + 0.012*"train" + 0.012*"skills" + 0.009*"diploma" + 0.009*"money" + 0.009*"degree" + 0.009*"time" + 0.007*"singaporean" + 0.007*"low" + 0.007*"pap"

Label: Retrenchment
Topic 0: 0.017*"pap" + 0.017*"budget" + 0.013*"future" + 0.009*"credits" + 0.009*"skills" + 0.009*"courses" + 0.009*"train" + 0.009*"skillsfuture" + 0.009*"skill" + 0.009*"lim"
Topic 1: 0.019*"budget" + 0.017*"job" + 0.014*"year" + 0.011*"better" + 0.011*"going" + 0.008*"old" + 0.008*"upgrade" + 0.008*"talk" + 0.008*"money" + 0.008*"olds"
Topic 2: 0.013*"reserves" + 0.013*"need" + 0.013*"current" + 0.010*"long" + 0.010*"singapore" + 0.007*"employ" + 0.007*"old" + 0.007*"good" + 0.007*"look" + 0.007*"living"
Topic 3: 0.018*"course" + 0.015*"money" + 0.011*"tsk" + 0.011*"wong" + 0.011*"skills" + 0.008*"think" + 0.008*"need" + 0.008*"people" + 0.008*"ask" + 0.008*"mr"
Topic 4: 0.016*"courses" + 0.011*"budget" + 0.011*"governments" + 0.011*"years" + 0.011*"singaporeans" + 0.011*"ones" + 0.006*"singapore" + 0.006*"election" + 0.006*"dpm" + 0.006*"psp"

```
Topic 5: 0.032*"job" + 0.016*"old" + 0.012*"diploma" + 0.012*"years" + 0.008*"singap
oreans" + 0.008*"low" + 0.008*"package" + 0.008*"poa" + 0.008*"bms" + 0.008*"degree"
Topic 6: 0.015*"party" + 0.011*"reserves" + 0.011*"future" + 0.011*"money" + 0.011
*"pap" + 0.011*"gov" + 0.011*"packet" + 0.010*"red" + 0.006*"skills" + 0.006*"using"
```

Step 5: Results and Discussion

Purpose:

To present and interpret the results from the LDA topic modeling, highlighting key findings and insights derived from the analysis.

Goal:

Provide a comprehensive understanding of the topics discussed within each dataset and label, and draw meaningful conclusions from the data.

Objective

Provide a comprehensive summary of the key themes, insights, and implications derived from the LDA topic modeling results on the combined social media dataset.

Key Themes Identified

1. Financial Concerns

CPF and Retirement Planning

- Frequent discussions on CPF accounts (e.g., OA, SA, RA) and their management.
- Concerns about financial security in retirement.
- **Keywords:** cpf, sa, ra, money, account, withdraw, retirement.

Cost of Living

- Topics highlighting the rising cost of living, GST increases, and financial burdens on Singaporeans.
- **Keywords:** gst, increase, cost, living, singaporeans, pay.

2. Political Sentiment

PAP and Government Policies

- Sentiment towards the PAP party and its effectiveness.
- Mixed views on the government's budget and financial policies.

- **Keywords:** pap, party, responsible, effective, policies, budget, gst.

3. Social and Economic Issues

Job Market and Employment

- Discussions on job availability, employment conditions, and government support for job creation.
- **Keywords:** job, singapore, employment, make, hope, future.

Income Inequality

- Concerns about wealth disparity, the rich vs. poor, and social stratification.
- **Keywords:** rich, poor, income, budget, life, serving.

4. Public Services and Vouchers

CDC Vouchers and Financial Assistance

- Conversations around the usage and effectiveness of CDC vouchers and other financial aids.
- **Keywords:** cdc, voucher, use, help, support.

Medisave and Healthcare

- Discussions on healthcare costs, Medisave, and access to medical services.
- **Keywords:** medisave, healthcare, cost, medical.

5. Quality of Life

General Well-being

- Topics on the overall quality of life, happiness, and public satisfaction.
- **Keywords:** life, happiness, caring, effective, fantastic.

Community and Social Cohesion

- Conversations about social harmony, community support, and national identity.
- **Keywords:** singaporeans, community, support, responsible.

Discussion

Financial Concerns

CPF and Retirement Planning

- The heavy emphasis on CPF-related discussions suggests a strong public interest in financial planning and retirement security. Concerns about managing CPF accounts and ensuring sufficient funds for retirement are prevalent.
- **Implications:** Financial advisors and policymakers should focus on improving communication about CPF policies and providing more resources for retirement planning.

Cost of Living

- Rising costs and GST increases are significant pain points for many individuals. These discussions reflect worries about affordability and economic stability.
- **Implications:** Policymakers might need to consider strategies to mitigate the impact of rising costs and ensure financial support for those affected.

Political Sentiment

PAP and Government Policies

- The mixed sentiment towards the PAP party indicates both support for and criticism of government actions. Discussions on policies, effectiveness, and governance are central themes.
- **Implications:** The government could benefit from addressing public concerns transparently and engaging with citizens to understand their needs better.

Social and Economic Issues

Job Market and Employment

- Employment conditions and job availability are recurring topics, indicating public concern about job security and opportunities.
- **Implications:** Government initiatives to create jobs and support employment could be crucial in addressing these concerns. Emphasis on upskilling and vocational training might also be beneficial.

Income Inequality

- The disparity between the rich and poor and its social implications are significant discussion points.
- **Implications:** Addressing income inequality through policies promoting fair wages and social welfare programs could help reduce economic disparities.

Public Services and Vouchers

CDC Vouchers and Financial Assistance

- Conversations on financial aids like CDC vouchers highlight their importance to the public.
- **Implications:** Ensuring the effective distribution and utilization of these aids can help alleviate financial burdens on lower-income groups.

Medisave and Healthcare

- Healthcare costs and access to medical services are important themes, with discussions around Medisave being particularly notable.
- **Implications:** Improving healthcare affordability and expanding access to medical services can address these public concerns.

Quality of Life

General Well-being

- Discussions about happiness and overall quality of life suggest that financial stability, social cohesion, and effective governance are key to public satisfaction.
- **Implications:** Policies aimed at improving quality of life, such as community support programs and mental health initiatives, could have a positive impact.

Community and Social Cohesion

- The importance of social harmony and national identity is evident in discussions about community and support.
- **Implications:** Promoting social cohesion through community-building activities and inclusive policies can strengthen national identity and unity.

Conclusion

The LDA topic modeling has identified several key themes in the combined social media dataset, offering valuable insights into public concerns and sentiments. These themes encompass financial planning, political sentiment, social and economic issues, public services, and quality of life.