*# This Python 3 environment comes with many helpful analytics libraries installed*

*# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python*

*# For example, here's several helpful packages to load*

import numpy as np *# linear algebra*

import pandas as pd *# data processing, CSV file I/O (e.g. pd.read\_csv)*

*# Input data files are available in the read-only "../input/" directory*

*# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory*

import os

for dirname, \_, filenames **in** os.walk('/kaggle/input'):

for filename **in** filenames:

print(os.path.join(dirname, filename))

*# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a version using "Save & Run All"*

*# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session*

/kaggle/input/fake-and-real-news-dataset/True.csv

/kaggle/input/fake-and-real-news-dataset/Fake.csv

In [2]:

!pip install gensim # Gensim is an open-source library for unsupervised topic modeling and natural language processing

import nltk

nltk.download('punkt')

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from wordcloud import WordCloud, STOPWORDS

import nltk

import re

from nltk.corpus import stopwords

import seaborn as sns

import gensim

from gensim.utils import simple\_preprocess

from gensim.parsing.preprocessing import STOPWORDS

import plotly.express as px

from sklearn.model\_selection import train\_test\_split

from sklearn.feature\_extraction.text import CountVectorizer

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import roc\_auc\_score

from sklearn.metrics import confusion\_matrix

Requirement already satisfied: gensim in /opt/conda/lib/python3.10/site-packages (4.3.2)

Requirement already satisfied: numpy>=1.18.5 in /opt/conda/lib/python3.10/site-packages (from gensim) (1.23.5)

Requirement already satisfied: scipy>=1.7.0 in /opt/conda/lib/python3.10/site-packages (from gensim) (1.11.2)

Requirement already satisfied: smart-open>=1.8.1 in /opt/conda/lib/python3.10/site-packages (from gensim) (6.3.0)

[nltk\_data] Downloading package punkt to /usr/share/nltk\_data...

[nltk\_data] Package punkt is already up-to-date!

**Import the data & Clean ups**

In [3]:

*#importing data*

fake\_data = pd.read\_csv('/kaggle/input/fake-and-real-news-dataset/Fake.csv')

print("fake\_data",fake\_data.shape)

true\_data= pd.read\_csv('/kaggle/input/fake-and-real-news-dataset/True.csv')

print("true\_data",true\_data.shape)

fake\_data (23481, 4)

true\_data (21417, 4) stop\_words = stopwords.words('english')

stop\_words.extend(['from', 'subject', 're', 'edu', 'use'])

def preprocess(text):

result = []

for token **in** gensim.utils.simple\_preprocess(text):

if token **not** **in** gensim.parsing.preprocessing.STOPWORDS **and** len(token) > 2 **and** token **not** **in** stop\_words:

result.append(token)

return result

In [9]:

*# Transforming the unmatching subjects*

In [4]:

fake\_data.head(5)

sub\_tf\_df=df.groupby('target').apply(lambda x:x['title'].count()).reset\_index(name='Counts')

sub\_tf\_df.target.replace({0:'False',1:'True'},inplace=True)

fig = px.bar(sub\_tf\_df, x="target", y="Counts",

color='Counts', barmode='group',

height=350)

fig.show()