Fake News Detection using NLP

PHASE III PROJECT SUBMISSION

TEAM MEMBERS

THIRUMALAI B   821021104049
TAMILSELVAN M   821021104048
KISHORE G M   821021104027
NITHISHKUMAR S   821021104035

Problem Statement:

Begin building the fake news detection model by loading and preprocessing the dataset. Load the fake news dataset and preprocess the textual data.

Data Cleaning:

Data cleaning is a process of removing inconsistencies in the dataset and incorrect values .It also in involves handling missing values either by removing them or assigning them average values. It helps to improve the efficiency of the model.

In the first step, we will only remove the unnecessary data points from the dataset which does not helps in improving the model performance.

Initially we import the necessary packages for our data cleaning process and also in the future purposes,

```
[ ]  import numpy as np
     import pandas as pd
     import matplotlib.pyplot as plt
     import seaborn as sns
     import nltk
     import re
     from wordcloud import WordCloud
     from tensorflow.keras.preprocessing.text import Tokenizer
     from tensorflow.keras.preprocessing.sequence import pad_sequences
     from tensorflow.keras.models import Sequential
     from tensorflow.keras.layers import Dense, Embedding, LSTM, Conv1D, MaxPool1D
     from sklearn.model_selection import train_test_split
     from sklearn.metrics import classification_report, accuracy_score
     import numpy as np
     import pandas as pd
```

we use these packages in various stages of our cleaning process and also in the future in which we need to build models.

Here, we read the .csv files of true and fake news and then explore the count values of their subjects

```
import os
for dirname, _, filenames in os.walk('/content/drive/MyDrive/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

/content/drive/MyDrive/input/True.csv
/content/drive/MyDrive/input/Fake.csv
```

```
fake_news = pd.read_csv('/content/drive/MyDrive/input/Fake.csv')
fake_news.head()
```

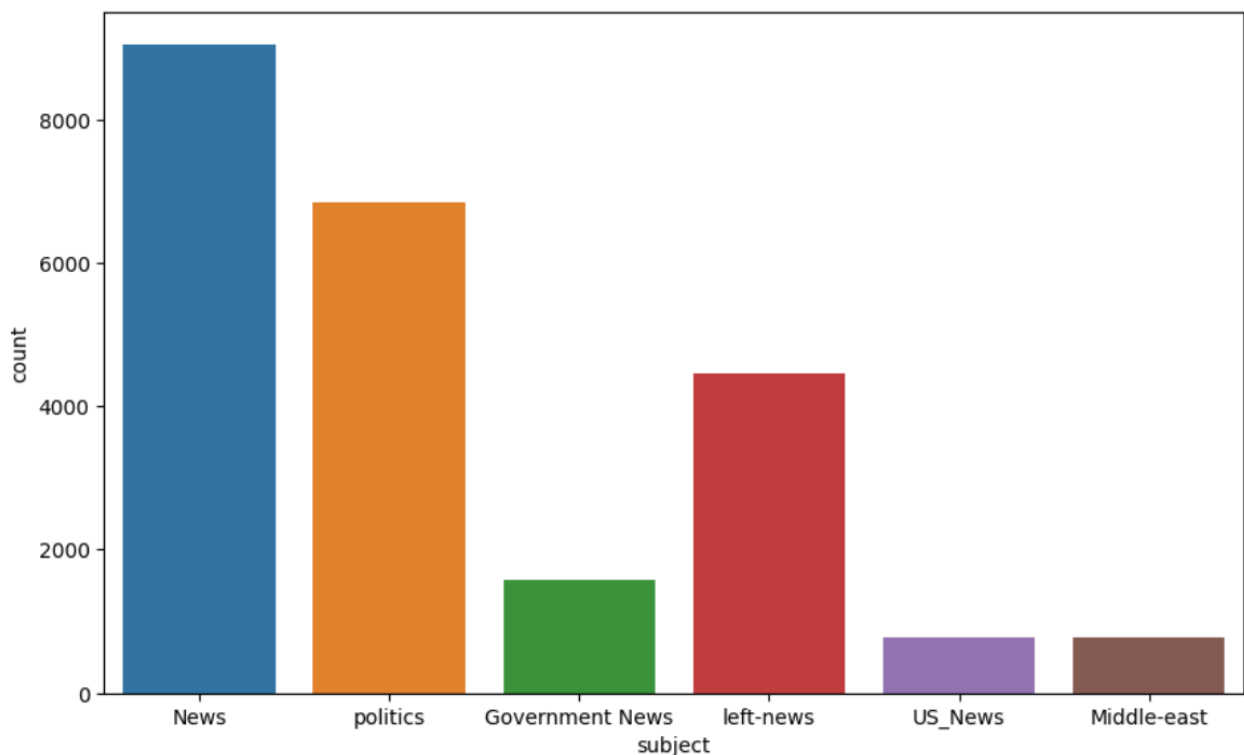|	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017

```
fake_news.columns
```

```
Index(['title', 'text', 'subject', 'date'], dtype='object')
```

```
fake_news['subject'].value_counts()
```

```
News              9050
politics          6841
left-news         4459
Government News   1570
US_News            783
Middle-east        778
Name: subject, dtype: int64
```

```
plt.figure(figsize=(10,6))
sns.countplot(x='subject',data=fake_news)
```
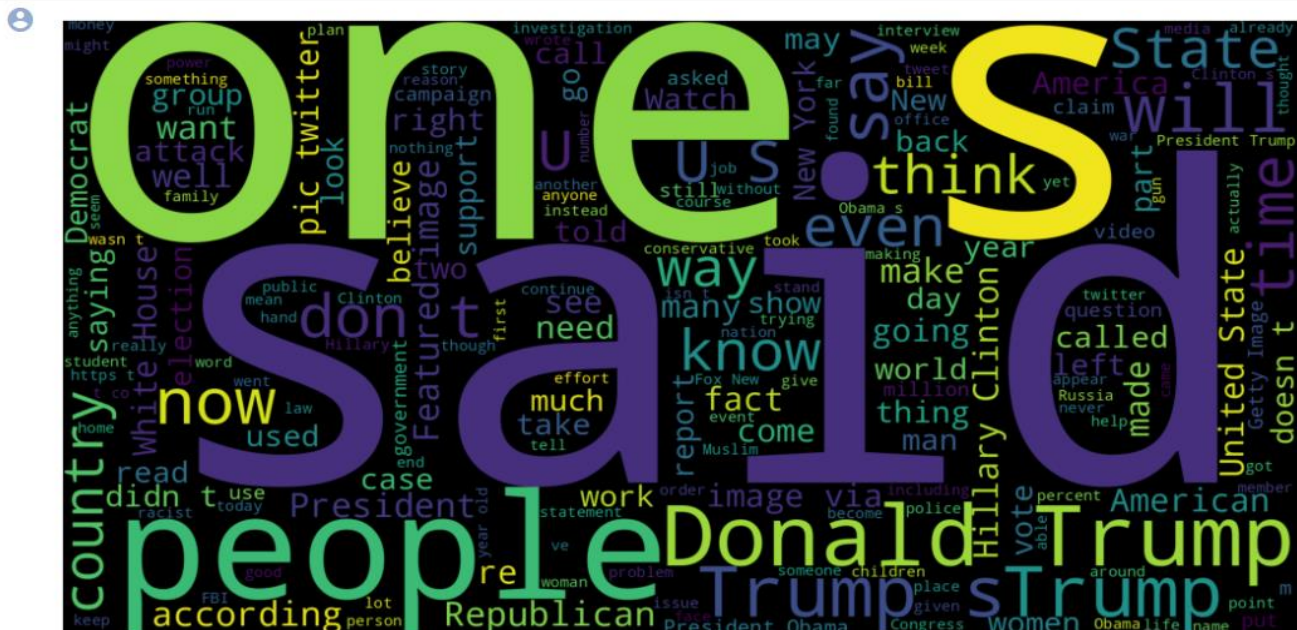
```
<Axes: xlabel='subject', ylabel='count'>
```

Here , we have used wordcloud to see that which word has mostly used for the fake news. By seeing that we can make a conclusion that which topic(about a person, event or anything) is mostly contains fake news).We also do the same for true news.

Word Cloud for Fake News:

```
%%time
wordcloud = WordCloud(width=1920, height=1000).generate(text)
fig = plt.figure(figsize=(10,10))
plt.imshow(wordcloud)
plt.axis('off')
plt.tight_layout(pad=0)
plt.show()
```



```
CPU times: user 43 s, sys: 3.33 s, total: 46.4 s
Wall time: 50.4 s
```
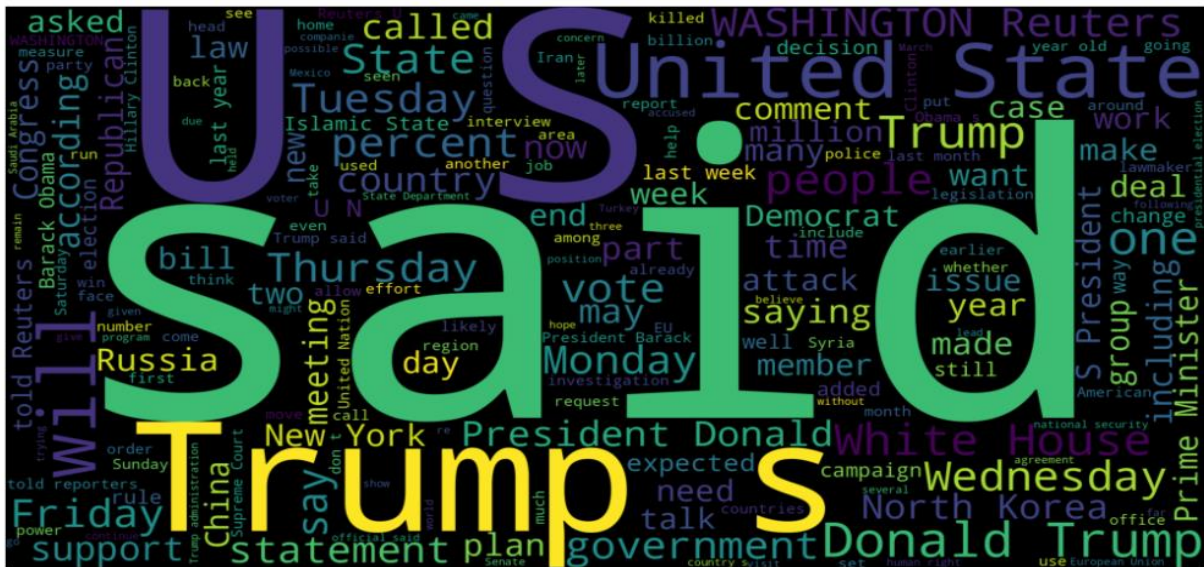
Word cloud for True News:

```
real_news = pd.read_csv('/content/drive/MyDrive/input/True.csv')
real_news.head()
```

	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017

```
[ ] text = ' '.join(real_news['text'].tolist())
```

```
[ ] %%time
    wordcloud = WordCloud(width=1920, height=1000).generate(text)
    fig = plt.figure(figsize=(10,10))
    plt.imshow(wordcloud)
    plt.axis('off')
    plt.tight_layout(pad=0)
    plt.show()
```



```
CPU times: user 38.5 s, sys: 2.83 s, total: 41.3 s
Wall time: 41.4 s
```

```
[ ] real_news.sample(5)
```

	title	text	subject	date
5933	Peru and Colombia vow to stand with Mexico aft...	LIMA (Reuters) - Peru and Colombia vowed to st...	politicsNews	January 27, 2017
15390	North Korean embassy official in focus at Kim ...	KUALA LUMPUR (Reuters) - Three men wanted for ...	worldnews	November 8, 2017
6088	Trump's exit from Pacific trade deal opens doo...	BERLIN (Reuters) - Germany would take advantag...	politicsNews	January 23, 2017
8915	Albanian town backs Clinton with bronze bust	SARANDE, Albania (Reuters) - Whatever the outc...	politicsNews	June 30, 2016
1319	House Republicans to take up disaster funding ...	WASHINGTON (Reuters) - U.S. House of Represent...	politicsNews	October 11, 2017

Let's create a list of news lists in real_news.csv with unknown publishers by using the following code snippets

```
[ ] unknown_pubishers = []
    for index, row in enumerate(real_news.text.values):
        try:
            record = row.split(' - ', maxsplit=1)
            record[1]

            assert(len(record[0])<260)
        except:
            unknown_pubishers.append(index)
```

```
[ ]  len(unknown_pubishers)

     35
```

```
     real_news.iloc[unknown_pubishers].text
```

```
2922      The following statements were posted to the ve...
3488      The White House on Wednesday disclosed a group...
3782      The following statements were posted to the ve...
4358      Neil Gorsuch, President Donald Trump's appoint...
4465      WASHINGTON The clock began running out this we...
5290      The following statements were posted to the ve...
5379      The following statements were posted to the ve...
5412      The following statements were posted to the ve...
5504      The following statements were posted to the ve...
5538      The following statements were posted to the ve...
5588      The following statements were posted to the ve...
5593      The following statements were posted to the ve...
5761      The following bullet points are from the U.S. ...
5784      Federal appeals court judge Neil Gorsuch, the ...
6026      The following bullet points are from the U.S. ...
6184      The following bullet points are from the U.S. ...
6660      Republican members of Congress are complaining...
6823      Over the course of the U.S. presidential campa...
7922      After going through a week reminiscent of Napo...
8194      The following timeline charts the origin and s...
8195      Global health officials are racing to better u...
8247      U.S. President Barack Obama visited a street m...
8465      ALGONAC, MICH.—Parker Fox drifted out of the D...
8481      Global health officials are racing to better u...
8482      The following timeline charts the origin and s...
8505      Global health officials are racing to better u...
8506      The following timeline charts the origin and s...
8771      In a speech weighted with America's complicate...
8970
9008      The following timeline charts the origin and s...
9009      Global health officials are racing to better u...
9307      It's the near future, and North Korea's regime...
9618      GOP leaders have unleashed a stunning level of...
9737      Caitlyn Jenner posted a video on Wednesday (Ap...
10479     The Democratic and Republican nominees for the...
Name: text, dtype: object
```

```python
publisher =[]
tmp_text = []

for index, row in enumerate(real_news.text.values):
    if index in unknown_pubishers:
        tmp_text.append(row)
        publisher.append('Unknown')

    else:
        record = row.split('-', maxsplit=1)
        publisher.append(record[0].strip())
        tmp_text.append(record[1].strip())
```

```
[ ]  real_news['publisher'] = publisher
     real_news['text'] = tmp_text
```

```
[ ]  real_news.head()
```

```
[ ]  real_news.head()
```

	title	text	subject	date	publisher
0	As U.S. budget fight looms, Republicans flip t...	The head of a conservative Republican faction ...	politicsNews	December 31, 2017	WASHINGTON (Reuters)
1	U.S. military to accept transgender recruits o...	Transgender people will be allowed for the fir...	politicsNews	December 29, 2017	WASHINGTON (Reuters)
2	Senior U.S. Republican senator: 'Let Mr. Muell...	The special counsel investigation of links bet...	politicsNews	December 31, 2017	WASHINGTON (Reuters)
3	FBI Russia probe helped by Australian diplomat...	Trump campaign adviser George Papadopoulos tol...	politicsNews	December 30, 2017	WASHINGTON (Reuters)
4	Trump wants Postal Service to charge 'much mor...	President Donald Trump called on the U.S. Post...	politicsNews	December 29, 2017	SEATTLE/WASHINGTON (Reuters)

```
[ ]  real_news.shape
```

```
(21417, 5)
```

```
[ ]  empty_fake_index = [index for index,text in enumerate(fake_news.text.tolist()) if str(text).strip()==""]
```

```
[ ]  fake_news.iloc[empty_fake_index]
```

	title	text	subject	date
10923	TAKE OUR POLL: Who Do You Think President Trum...		politics	May 10, 2017
11041	Joe Scarborough BERATES Mika Brzezinski Over "...		politics	Apr 26, 2017
11190	WATCH TUCKER CARLSON Scorch Sanctuary City May...		politics	Apr 6, 2017
11225	MAYOR OF SANCTUARY CITY: Trump Trying To Make ...		politics	Apr 2, 2017
11236	SHOCKER: Public School Turns Computer Lab Into...		politics	Apr 1, 2017
...	...	...	...	...
21816	BALTIMORE BURNS: MARYLAND GOVERNOR BRINGS IN N...		left-news	Apr 27, 2015
21826	FULL VIDEO: THE BLOCKBUSTER INVESTIGATION INTO...		left-news	Apr 25, 2015
21827	(VIDEO) HILLARY CLINTON: RELIGIOUS BELIEFS MUS...		left-news	Apr 25, 2015
21857	(VIDEO)ICE PROTECTING OBAMA: WON'T RELEASE NAM...		left-news	Apr 14, 2015
21873	(VIDEO) HYSTERICAL SNL TAKE ON HILLARY'S ANNOU...		left-news	Apr 12, 2015

630 rows × 4 columns

```
[ ]  real_news['text'] = real_news['title']+" "+ real_news['text']
```

```
[ ]  fake_news['text'] = fake_news['title']+" "+ fake_news['text']
```

```
[ ]  real_news['text'] = real_news['text'].apply(lambda x: str(x).lower())
     fake_news['text'] = fake_news['text'].apply(lambda x: str(x).lower())
```

```
[ ]  real_news['class']=1
     fake_news['class']=0
```

```
[ ]  real_news.columns
```

```
Index(['title', 'text', 'subject', 'date', 'publisher', 'class'], dtype='object')
```

```
▶  real = real_news[['text','class']]
   fake = fake_news[['text','class']]
```

```
▶  data = real.append(fake, ignore_index=True)
   data.head()
```

```
<ipython-input-33-8770c2a2a545>:1: FutureWarning: The frame.append method is deprecated and will be removed from pandas in a future version. Use pandas.concat instead.
  data = real.append(fake, ignore_index=True)
```

	text	class
0	as u.s. budget fight looms, republicans flip t...	1
1	u.s. military to accept transgender recruits o...	1
2	senior u.s. republican senator: 'let mr. muell...	1
3	fbi russia probe helped by australian diplomat...	1
4	trump wants postal service to charge 'much mor...	1

```
!pip install spacy==2.2.3
!python -m spacy download en_core_web_sm
!pip install beautifulsoup4==4.9.1!
!pip install textblob==0.15.3
!pip install git+https://github.com/laxmimerit/preprocess_kgptalkie.git --upgrade --force-reinstall
```

note: This error originates from a subprocess, and is likely not a problem with pip.
2023-10-16 16:41:23.058566: W tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not find TensorRT
Collecting en-core-web-sm==3.6.0
  Downloading https://github.com/explosion/spacy-models/releases/download/en_core_web_sm-3.6.0/en_core_web_sm-3.6.0-py3-none-any.whl (12.8 MB)
                                          12.8/12.8 MB 21.6 MB/s eta 0:00:00
Requirement already satisfied: spacy<3.7.0,>=3.6.0 in /usr/local/lib/python3.10/dist-packages (from en-core-web-sm==3.6.0) (3.6.1)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (3.0.12)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (1.0.5)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (1.0.10)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.0.8)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (3.0.9)
Requirement already satisfied: thinc<8.2.0,>=8.1.8 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (8.1.12)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (1.1.2)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.4.8)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.0.10)
Requirement already satisfied: typer<0.10.0,>=0.3.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (0.9.0)
Requirement already satisfied: pathy>=0.10.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (0.10.2)
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (6.4.0)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (4.66.1)
Requirement already satisfied: numpy>=1.15.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (1.23.5)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.31.0)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (1.10.13)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (3.1.2)
Requirement already satisfied: setuptools in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (67.7.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (23.2)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (3.3.0)
Requirement already satisfied: typing-extensions>=4.2.0 in /usr/local/lib/python3.10/dist-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (4.5.0)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (3.3.0)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.0.6)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2023.7.22)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in /usr/local/lib/python3.10/dist-packages (from thinc<8.2.0,>=8.1.8->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (0.7.11)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in /usr/local/lib/python3.10/dist-packages (from thinc<8.2.0,>=8.1.8->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (0.1.3)
Requirement already satisfied: click<9.0.0,>=7.1.1 in /usr/local/lib/python3.10/dist-packages (from typer<0.10.0,>=0.3.0->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (8.1.7)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-packages (from jinja2->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.1.3)

```python
import preprocess_kgptalkie as ps
```

```python
data['text'] = data['text'].apply(lambda x: ps.remove_special_chars(x))
```