Assignment - 2

Task 1:

This dataset contains a collection of Reddit posts, including columns such as "title" (post title), "text" (post content), "score" (post score), and "upvote_ratio" (upvote ratio). Additionally, it includes empty or unnamed columns, which may need cleaning.

	title	text	score	upvote_ratio	upvotes	Unnamed: 5	Unnamed: 6
0	Looking for married Muslim men who have hijabi...	Don't message me if you can't live verify. Too...	1	1	1	NaN	NaN
1	Share your istikhara success stories. I need s...	Salaam everyone. I'm a F currently going throu...	1	1	1	NaN	NaN
2	Fate?	In the Qur'an, I saw verses in these cases tha...	1	1	1	NaN	NaN
3	Good Thrift shop find. Highly reccomend	Holocaust book about family of Jewish Hungaria...	1	1	1	NaN	NaN
4	Wearing my kippah with tattoos	Shalom friends!\n\nI'm a Baal teshuva with man...	1	1	1	NaN	NaN

The dataset contains 92,815 entries with seven columns, with the majority of the entries being non-null objects. The last two columns, "Unnamed: 5" and "Unnamed: 6," have only one non-null entry, suggesting they may not hold significant information and could be candidates for removal during data cleaning.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 92815 entries, 0 to 92814
Data columns (total 7 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   title         92671 non-null  object
 1   text          92670 non-null  object
 2   score         92659 non-null  object
 3   upvote_ratio  92658 non-null  object
 4   upvotes       92656 non-null  object
 5   Unnamed: 5    1 non-null      object
 6   Unnamed: 6    1 non-null      object
dtypes: object(7)
memory usage: 5.0+ MB
```

We identified and extracted rows from the dataset where the "text" column contains URLs.The result shows that 17,485 rows contain URLs, highlighting their presence within the text column. The URLs from the texts are then removed.

```
Sentences with URLs:
                                                    text
9      i found an old bbc documentary show titled "co...
26     /r/worldnews\n   \n https://www.reuters.com/wo...
31           https://www.youtube.com/watch?v=v3ga7skxino
50     this is an automated weekly thread\n\nfeel fre...
61     /r/worldnews\n   \n https://www.forbes.com/sit...
...                                                  ...
92653  israeli authorities have declined to release a...
92656  /r/worldnews\n   \n https://www.businessinside...
92657  it is important for those new to the palestini...
92662  [https://bloggingtheology.net/2016/05/06/chris...
92669  my previous post was removed by a mod because ...

[17485 rows x 1 columns]
```

We identified rows containing email addresses in the "text" column using a regular expression pattern. The email addresses are then removed.

```
Sentences with email addresses:
                                                    text
7798   i'm running a speaker event on monday evening,...
14458   hi i'm a disabled  dad if 2, london uk\nmy wi...
19960  i am excited that laasok - the new liberal bei...
22489  i need a couple of volunteers for tomorrow.  s...
25704  salamalaikum to everyone. i want to help out s...
25962  hi all. i'm a reform rabbi, and founder of laa...
25966  hello! we are a group of undergraduates lookin...
26537  hello, my name is veronica, and i am a doctora...
26540  hello, my name is veronica, and i am a doctora...
26848  \n\nanasya's donation link: anastasiyaparaskev...
40437  this is their official [post on fb:](\n\ndear ...
41226  on 12/10/2023 on lbc at around 7:00pm-8:00pm t...
42619  binance, the world's largest cryptocurrency ex...
43429  \n\tplease use the sharing tools found via the...
48964    \nplease use the sharing tools found via the...
55899  reform judaism is often referred to as "non-ha...
61884  to all those who reside in canada, please boos...
62578  i would post to /r/canada and /r/onguardforthe...
63400  “humanrightswatch - we are documenting censors...
63991  has its hands ties since israel and the usa wo...
64348  the washington post is interested in hearing f...
65280  at faith to faithless (a service run by the ch...
65561  hi friends. a lot of people are struggling rig...
65746  any pro israel folks want to help get another ...
65851  [screenshot of the tweet where he offers the r...
66796  if you’re in biglaw and spend any time at all ...
71555  hello!\n\ni am part of the peace and justice p...
72488  i should profit of this but the war is on the ...
75909  his name is brian kaplan, an employee at @ever...
80015   hello r/askmiddleeast community,\n\ni hope th...
80016   hello r/askmiddleeast community,\n\ni hope th...
83493  hello! my name is adriana, i am a senior psych...
```

We identified rows containing hashtags in the "text" column using a regular expression pattern. It found 4,728 rows where hashtags are present. The hashtags are then removed.

```
Sentences with hashtags:
                                                    text
5      slightly embarrassing question but also someth...
6      i found this subreddit not too long ago but up...
50     this is an automated weekly thread\n\nfeel fre...
93     \* you live in the west and are casually walki...
96                                                #name?
...                                                  ...
92589                      &#x200b;\n\n\n\n&#x200b;\n\n
92595                      &#x200b;\n\n\n\n&#x200b;\n\n
92618  hi all,\n\ni wrote my congressman today (2nd t...
92653  israeli authorities have declined to release a...
92657  it is important for those new to the palestini...

[4728 rows x 1 columns]
```

Then we implemented a text preprocessing pipeline to clean and normalize textual data. This involved expanding contractions, removing special characters, numbers, and extra whitespace for uniformity. Stopwords were filtered out, and tokens were lemmatized based on their Part-of-Speech (POS) tags using WordNet.

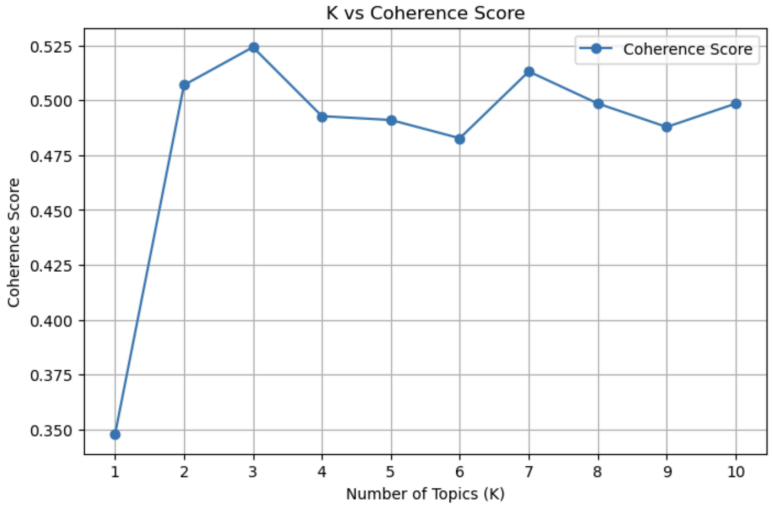
We created a WordCloud visualization to identify the most frequently occurring words in the cleaned text data. Common words such as "say," "think," "know," "one," and "even" are prominently displayed, reflecting their higher frequency and relevance in the dataset.

We performed topic modeling using the Latent Dirichlet Allocation (LDA) algorithm. First, the text was tokenized, and a dictionary of unique words was created using gensim. A corpus was generated by converting tokenized texts into a bag-of-words representation. LDA models with varying numbers of topics (1 to 10) were built, and their coherence scores were calculated to evaluate the models.

```
coherence_scores
```

```
[0.3476693000754619,
 0.5070021711537325,
 0.5243005664142573,
 0.49282408971612557,
 0.4910690528118281,
 0.4827493229388195,
 0.513204288171699,
 0.49860038768039566,
 0.4878813142608602,
 0.4985342784745037]
```

The graph shows the coherence scores for the LDA model with different numbers of topics (K). The coherence score, which measures the interpretability of topics, peaks at K=3, indicating that the model with three topics provides the most meaningful grouping of the text data.



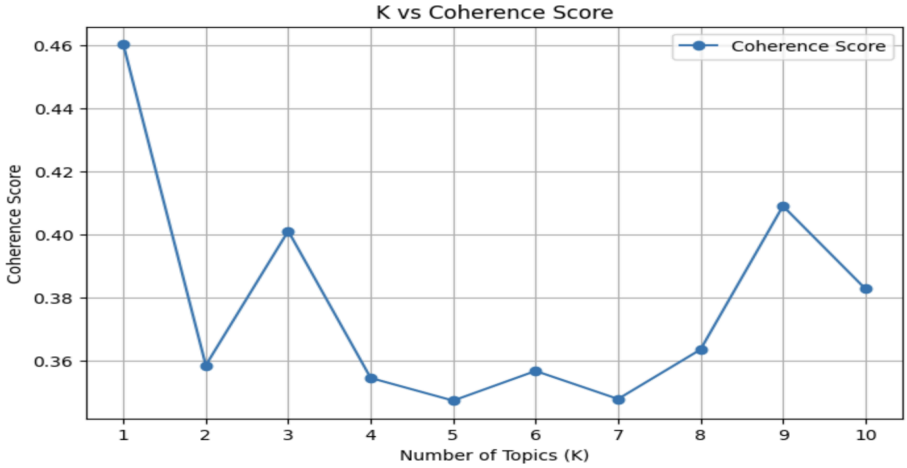
A WordCloud for LDA was plotted to visualize the most prominent words for each topic identified by the model.

Coherence scores were computed for BERTopic models with different numbers of topics to evaluate topic quality and interpretability.

```
coherence_scores
```

```
[0.46034903743790684,
 0.358379016708222,
 0.4010636159959726,
 0.3543446812353559,
 0.3471541813052466,
 0.3565785869485267,
 0.34767107695375615,
 0.36338051647907343,
 0.4090101126944182,
 0.3826890175611032]
```

The graph shows coherence scores for BERTopic models across different numbers of topics (K). The highest coherence score is observed at K=1.



This WordCloud for BERTopic displays the most significant words across topics identified by the model.

LDA's coherence score peaks at **K=3** with a value of approximately **0.525**, which suggests good interpretability of topics with three topics.
BERTopic's coherence score peaks at **K=1** with a value of approximately **0.46**, which is lower than LDA's optimal coherence score, indicating slightly less interpretable topics overall.
LDA provides more interpretable topics for this dataset because it aligns better with its structure and content. The dataset seems relatively well-preprocessed and tokenized, with topics having distinct separations that align well with LDA's word-based generative model.

Task 2:

We defined a function that identifies and extracts sentences from the text that mention specific entities, such as "hamas" or "israel." It uses a Named Entity Recognition (NER) pipeline to analyze the text and match entities from a predefined list. 17767 entries containing “hamas” and “israel” were extracted.

	entity_context
84	[case germany right author claim jewish year p...
91	[melachim ii ii king chapter ahaziah twentytwo...
102	[hello interest know jewish israeli think arab...
105	[may move israel near future take job offer je...
121	[expect mail israeli university outside israel...
...	...
92650	[see uptick amount comment accuse people suppo...
92653	[israeli authority decline release ahmad manas...
92657	[important new palestinian topic understand pa...
92661	[past two month leave frustrated upset native ...
92667	[every month politician west across almost par...

17767 rows × 1 columns

Example data point after NER: 'case germany right author claim jewish year publish opinion piece jewish life israel become sort public figure circle admit jewish tried excuse long essay publish major newspaper claim know real identity recently find become pretty clear probably lie know along though might

somehow suppressed regular german nazi grandparent fake jewish identity something happen fairly regularly germany still wonder thing outside germany bonus question likely jew name christian’

The data loaded is the valence-NRC-VAD-Lexicon, containing 19,971 rows with words and their corresponding valence scores (ranging from 0 to 1).

	word	score
0	generous	1.000
1	love	1.000
2	very positive	1.000
3	magnificent	1.000
4	happily	1.000
...	...	...
19966	disheartening	0.010
19967	mistreated	0.010
19968	toxic	0.008
19969	nightmare	0.005
19970	shit	0.000

19971 rows × 2 columns

The data contains 19,970 entries with two columns: "word" (object type) and "score" (float64 type)

```
<class 'pandas.core.frame.DataFrame'>
Index: 19970 entries, 0 to 19970
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   word    19970 non-null  object 
 1   score   19970 non-null  float64
dtypes: float64(1), object(1)
memory usage: 468.0+ KB
```

The results show that the average sentiment for "hamas" is 0.541, while for "israel," it is 0.556, reflecting the sentiment polarity of associated words in the text.

```
Average sentiment for 'hamas': 0.5412050473186111
Average sentiment for 'israel': 0.5556480280929228
```

This data is the NRC VAD Dominance Lexicon containing 19,971 rows with words and their corresponding dominance scores.

	word	score
0	powerful	0.991
1	leadership	0.983
2	success	0.981
3	govern	0.980
4	supreme	0.974
...	...	...
19966	vague	0.083
19967	penniless	0.083
19968	empty	0.081
19969	frail	0.069
19970	weak	0.045

19971 rows × 2 columns

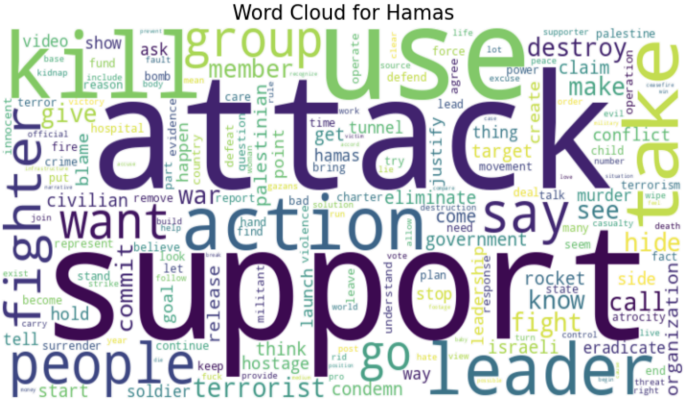
The data contains 19,970 entries with two columns: "word" (object type) and "score" (float64 type)

```
<class 'pandas.core.frame.DataFrame'>
Index: 19970 entries, 0 to 19970
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   word    19970 non-null  object
 1   score   19970 non-null  float64
dtypes: float64(1), object(1)
memory usage: 468.0+ KB
```

The results show that the average dominance for "Hamas" is **0.5837**, and for "Israel," it is **0.5788**, reflecting a similar level of dominance expressed in the text for both entities. This analysis combines dependency parsing with dominance scoring for entity-specific context.

```
Average Dominance for Hamas: 0.5836764968848978
Average Dominance for Israel: 0.5788739515677069
```

The word clouds display the most prominent words associated with "Hamas" and "Israel" in the analyzed text. For "Hamas," terms like "attack," "support," "fight," and "action" dominate, reflecting themes of conflict and military actions.



For "Israel," words like "defend," "start," "claim," and "kill" are prominent, also emphasizing themes of defense and conflict.

Both word clouds contain terms like "fight," "kill," "action," and "people," highlighting the shared context of conflict and violence involving both entities.

For **Hamas**, terms like "terrorist," "hide," "rocket," and "hostage" indicate an emphasis on perceived militant activities and tactics.

For **Israel**, terms like "defend," "occupy," "right," and "state" reflect a focus on self-defense, territorial claims, and governance.

Task 3:

We filter the dataset by retaining only rows where the score column contains valid numeric values.

	text	score
59434	reject peel commission idea partition palestin...	000 for five years and ruled that further immi...
59435	would function provisional governmentinexile a...	they were concerned over the heavy losses the...
59437	young jew sail palestine take mandatory palest...	initially under a different name. Lehi was a ...
59440	effectively end left hitherto unrivalled domin...	promoting a socially conservative and economi...
72394	etc propalestinians truly care much life gaza	why are they not calling for the surrender of...

We performed fine-tuning of a BERT-based model for regression tasks using a specific fold of training and testing data. The ClassificationModel is initialized with parameters for regression, batch size, learning rate, and sequence length, among others. Gradients for embeddings and the first eight encoder layers of BERT are frozen to reduce computational overhead.

```
Average MSE across 10 folds: 532.1835117871276
MSE Scores: [115.33005103033778, 380.6635658120066, 675.4687039945294, 246.68230745723713, 587.5802091136591, 248.9
6024680079054, 149.8327168485024, 619.8104285089831, 2003.2236024136998, 294.28328589153006]
```

The mse_scores list contains the individual MSE values for each fold, ranging from 115.33 to 2003.22, with significant variability between folds. The average MSE across all folds is 532.18, which serves as an overall metric to evaluate the model's performance. This variability indicates potential differences in fold characteristics or challenges in the dataset for certain splits.

References

1. https://www.geeksforgeeks.org/latent-dirichlet-allocation/
2. https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0
3. https://simpletransformers.ai/docs/regression/
4. https://github.com/elyesmanai/simpletransformerss