INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY
NEW DELHI

Department of Computer Science & Engineering

CSE 556 : Natural Language Processing (NLP)

Prof. Shad Akhtar

EDA Results

Anant Kumar Kaushal (2022067)
Anikait Agrawal (2022072)
Ansh Varshney (2022083)

Description of Procedure:

Training Dataset Analysis:

Initially, the training dataset is loaded into a Pandas DataFrame, and its structure is verified by displaying the dataset's shape, first few records, and column-wise information, including data types and non-null counts. A comprehensive statistical summary of all columns, covering both numerical and categorical data, is obtained using .describe(include='all'). Subsequently, exploratory data analysis is conducted by examining the distributions of key categorical features (csType, Suggest, Relevance, Aggresive, and Complexity). For each of these features, value counts are computed, and the distributions are visualized using count plots, enabling a clear assessment of class balance and representation. Additionally, the dataset is scrutinized for missing values across all columns to inform subsequent preprocessing decisions.

=== Train Data Head ===

	hatespeech	csType	counterspeech	Suggest	Relevance	Aggressive	Complexity	Comments	source	claim	...	hatespeechTa
0	Maybe the UN could talk to those asian and afr...	Informative	The us is the second most polluting country in...	3	4.0	2.0	3.0	NaN	Human	The UN should focus on Asian and African natio...	...	asian_p
1	Maybe the UN could talk to those asian and afr...	Questioning	Doesn't everyone on the planet have a responsi...		3.0	2.0	2.0	NaN	Human	The UN should focus on Asian and African natio...	...	asian_p
2	Maybe the UN could talk to those asian and afr...	Denouncing	The world would be a better place if people we...	1	1.0	2.0	2.0	NaN	Human	The UN should focus on Asian and African natio...	...	asian_p

```
Data columns (total 31 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   hatespeech                      9532 non-null   object
 1   csType                          9532 non-null   object
 2   counterspeech                   9532 non-null   object
 3   Suggest                         4108 non-null   object
 4   Relevance                       4100 non-null   float64
 5   Aggressive                      4104 non-null   float64
 6   Complexity                      4107 non-null   float64
 7   Comments                        0 non-null      float64
 8   source                          9532 non-null   object
 9   claim                           9520 non-null   object
 10  centralTopic                    9520 non-null   object
 11  speakerIntent                   9516 non-null   object
 12  targetGroup                     9516 non-null   object
 13  relevantPowerDynamics           9516 non-null   object
 14  hatespeechImplication           9516 non-null   object
 15  targetGroupEmotionalReaction    9516 non-null   object
 16  targetGroupCognitiveReaction    9516 non-null   object
 17  hatespeechOffensiveness         9532 non-null   object
 18  id                              9532 non-null   int64
 19  is_high_quality                 9532 non-null   object
 20  hs_id                           9532 non-null   int64
 21  hatespeechTarget                9532 non-null   object
 22  powerDynamics                   9532 non-null   object
 23  prompt_offensiveness            9532 non-null   object
 24  prompt_target_group             9532 non-null   object
 25  prompt_speaker_intent           9532 non-null   object
 26  prompt_power_dynamics           9532 non-null   object
 27  prompt_implication              9532 non-null   object
 28  prompt_emotional_reaction       9532 non-null   object
 29  prompt_cognitive_reaction       9532 non-null   object
 30  prompt_cs_generation            9532 non-null   object
```

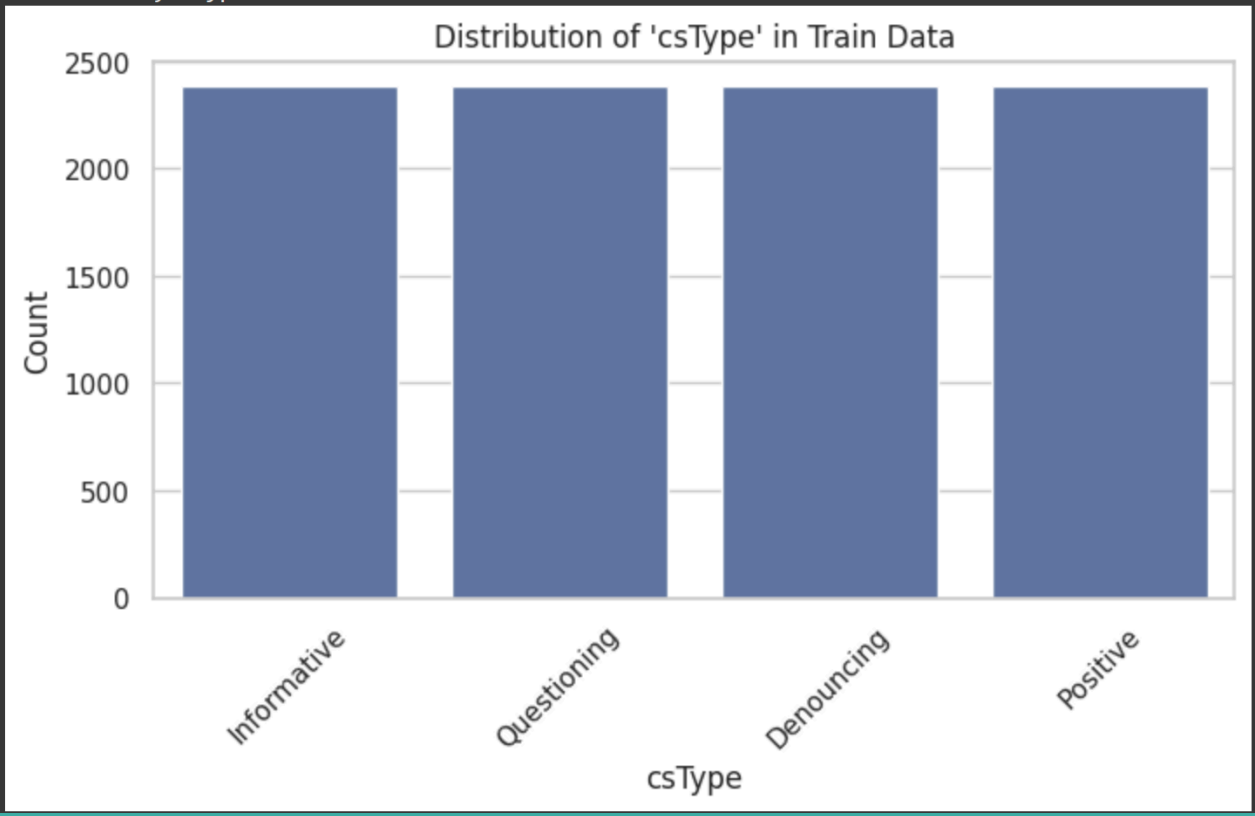
```
=== Train Data Summary (All Columns) ===
                                                hatespeech       csType  \
count                                                 9532         9532
unique                                                2383            4
top     The Holocaust never happened. Jews are an infe...  Informative
freq                                                     4         2383
mean                                                   NaN          NaN
std                                                    NaN          NaN
min                                                    NaN          NaN
25%                                                    NaN          NaN
50%                                                    NaN          NaN
75%                                                    NaN          NaN
max                                                    NaN          NaN

                                           counterspeech Suggest  \
count                                               9532    4108
unique                                              9529       4
top     It is the best solution to prevent conflicts w...       3
freq                                                   2    2718
mean                                                 NaN     NaN
std                                                  NaN     NaN
min                                                  NaN     NaN
25%                                                  NaN     NaN
50%                                                  NaN     NaN
75%                                                  NaN     NaN
max                                                  NaN     NaN
```

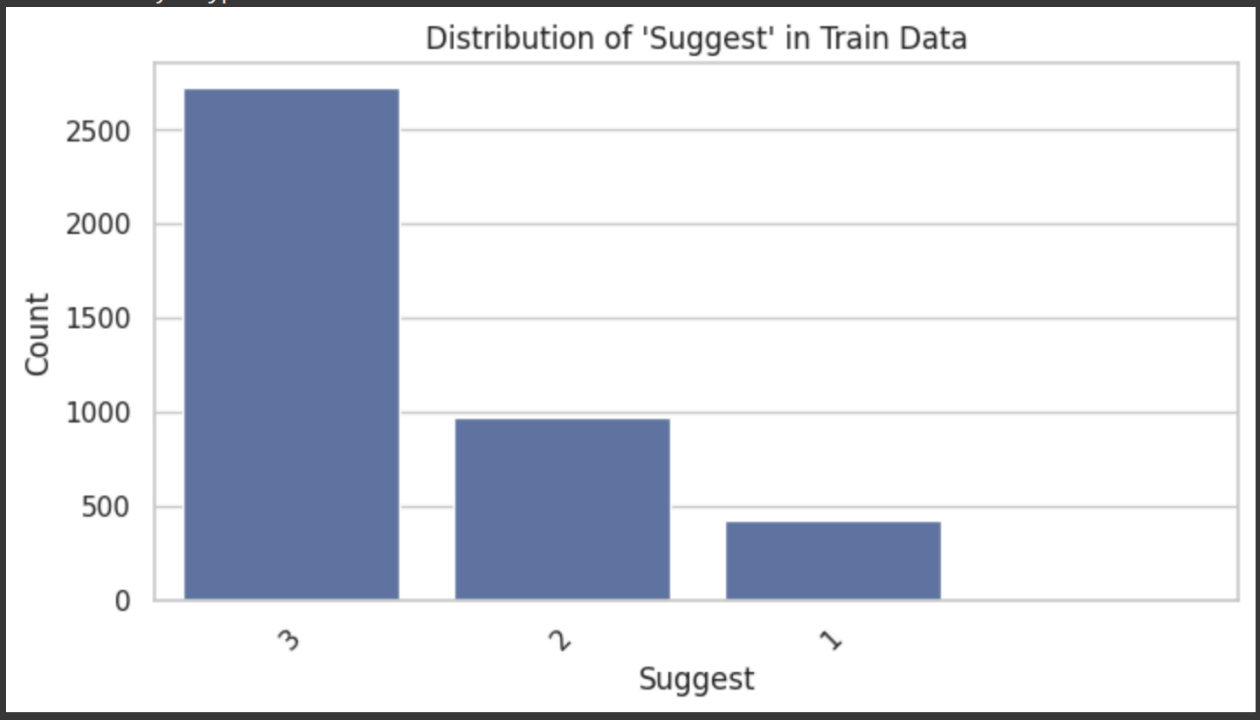
```
          Relevance   Aggressive   Complexity  Comments source  \
count   4100.000000  4104.000000  4107.000000       0.0   9532
unique          NaN          NaN          NaN       NaN      4
top             NaN          NaN          NaN       NaN  Human
freq            NaN          NaN          NaN       NaN   4116
mean       3.382683     2.279483     1.959338       NaN    NaN
std        1.067675     0.906880     0.637023       NaN    NaN
min        1.000000     1.000000     1.000000       NaN    NaN
25%        3.000000     2.000000     2.000000       NaN    NaN
50%        3.000000     2.000000     2.000000       NaN    NaN
75%        4.000000     3.000000     2.000000       NaN    NaN
max        5.000000     5.000000     5.000000       NaN    NaN

                    claim  ... hatespeechTarget powerDynamics  \
count                9520  ...            9532          9532
unique               2346  ...              15             8
top     Gays are perverts.  ...         muslims      religion
freq                   16  ...            1876          3076
mean                  NaN  ...             NaN           NaN
std                   NaN  ...             NaN           NaN
min                   NaN  ...             NaN           NaN
25%                   NaN  ...             NaN           NaN
50%                   NaN  ...             NaN           NaN
75%                   NaN  ...             NaN           NaN
max                   NaN  ...             NaN           NaN
```

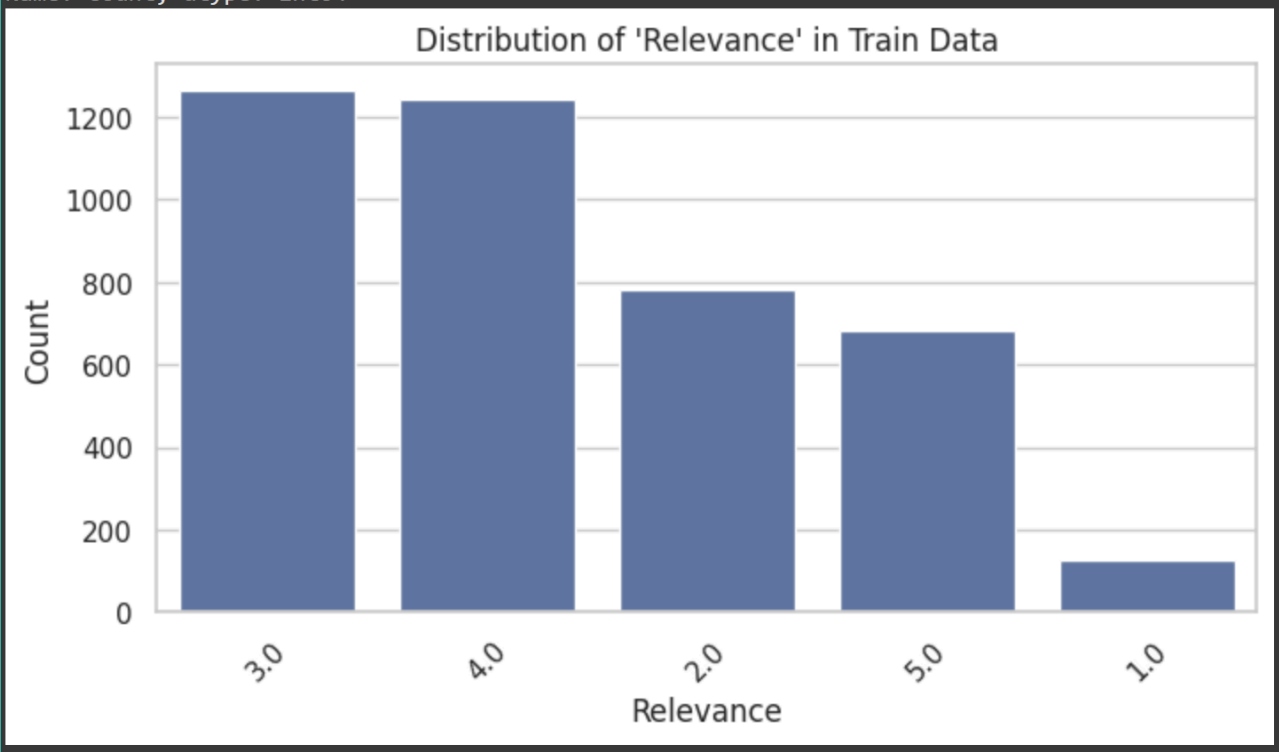
```
Value counts for 'csType' in Train Data:
csType
Informative     2383
Questioning     2383
Denouncing      2383
Positive        2383
Name: count, dtype: int64
```



```
Value counts for 'Suggest' in Train Data:
Suggest
3    2718
2     969
1     420
        1
Name: count, dtype: int64
```

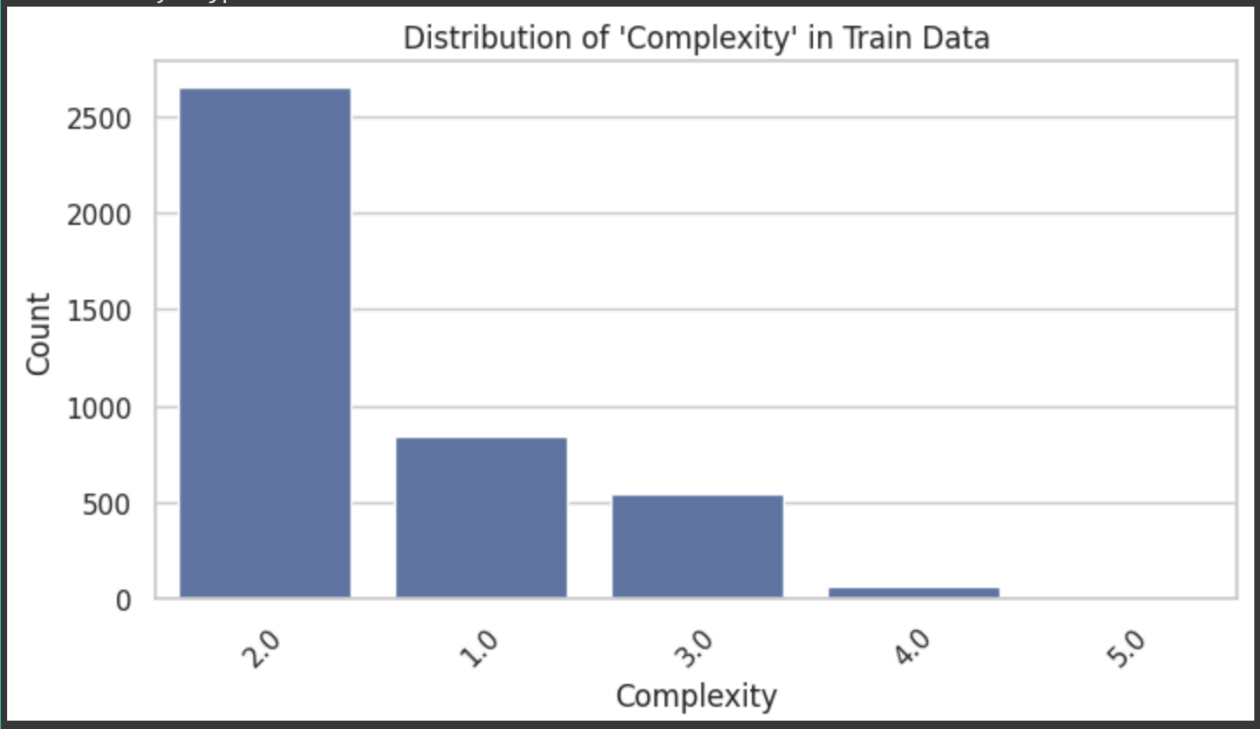


```
Value counts for 'Relevance' in Train Data:
Relevance
3.0    1263
4.0    1243
2.0     782
5.0     683
1.0     129
Name: count, dtype: int64
```



```
Column 'Aggresive' not found in Train Data.

Value counts for 'Complexity' in Train Data:
Complexity
2.0    2650
1.0     847
3.0     543
4.0      64
5.0       3
Name: count, dtype: int64
```



Validation Dataset Analysis:

The validation dataset undergoes an identical analytical procedure to ensure consistency between training and validation data processing. After loading and verifying the validation data's shape and first few records, column-specific data types and completeness are examined. A statistical summary for all columns is generated to offer detailed insights into the data distribution. Distributions for the key categorical features (`csType`, `Suggest`, `Relevance`, `Aggresive`, and `Complexity`) are analyzed through value counts and count plot visualizations, facilitating a direct comparison of data characteristics with the training set. Lastly, a thorough check for missing values in the validation dataset is performed to ensure data integrity and readiness for modeling and evaluation.

```
=== Validation Data Head ===
```

	index	hatespeech	csType	counterspeech	Suggest	Relevance	Aggressive	Complexity	Comments	source	...	hatespeechTarget
0	21	Women are basically childlike, they remain thi...	Questioning	How could you say something like this inspite ...	3.0	4.0	3.0	3.0	NaN	Human	...	women
1	50	Minorities vote to strip wealth and power from...	Denouncing	Aside from your statement having no factual ba...	3.0	4.0	3.0	2.0	NaN	Human	...	jews
2	141	We began to break down when we allowed women t...	Informative	Why is being 'soft' a problem? everyone is ent...	3.0	3.0	2.0	2.0	NaN	Human	...	women

```
Data columns (total 32 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   index                           1470 non-null   int64
 1   hatespeech                      1470 non-null   object
 2   csType                          1470 non-null   object
 3   counterspeech                   1470 non-null   object
 4   Suggest                         596 non-null    float64
 5   Relevance                       595 non-null    float64
 6   Aggressive                      596 non-null    float64
 7   Complexity                      596 non-null    float64
 8   Comments                        0 non-null      float64
 9   source                          1470 non-null   object
 10  claim                           1470 non-null   object
 11  centralTopic                    1470 non-null   object
 12  speakerIntent                   1470 non-null   object
 13  targetGroup                     1470 non-null   object
 14  relevantPowerDynamics           1470 non-null   object
 15  hatespeechImplication           1470 non-null   object
 16  targetGroupEmotionalReaction    1470 non-null   object
 17  targetGroupCognitiveReaction    1470 non-null   object
 18  hatespeechOffensiveness         1470 non-null   object
 19  id                              1470 non-null   int64
 20  is_high_quality                 1470 non-null   object
 21  hs_id                           1470 non-null   int64
 22  hatespeechTarget                1470 non-null   object
 23  powerDynamics                   1470 non-null   object
 24  prompt_offensiveness            1470 non-null   object
 25  prompt_target_group             1470 non-null   object
 26  prompt_speaker_intent           1470 non-null   object
 27  prompt_power_dynamics           1470 non-null   object
 28  prompt_implication              1470 non-null   object
 29  prompt_emotional_reaction       1470 non-null   object
 30  prompt_cognitive_reaction       1470 non-null   object
 31  prompt_cs_generation            1470 non-null   object
```

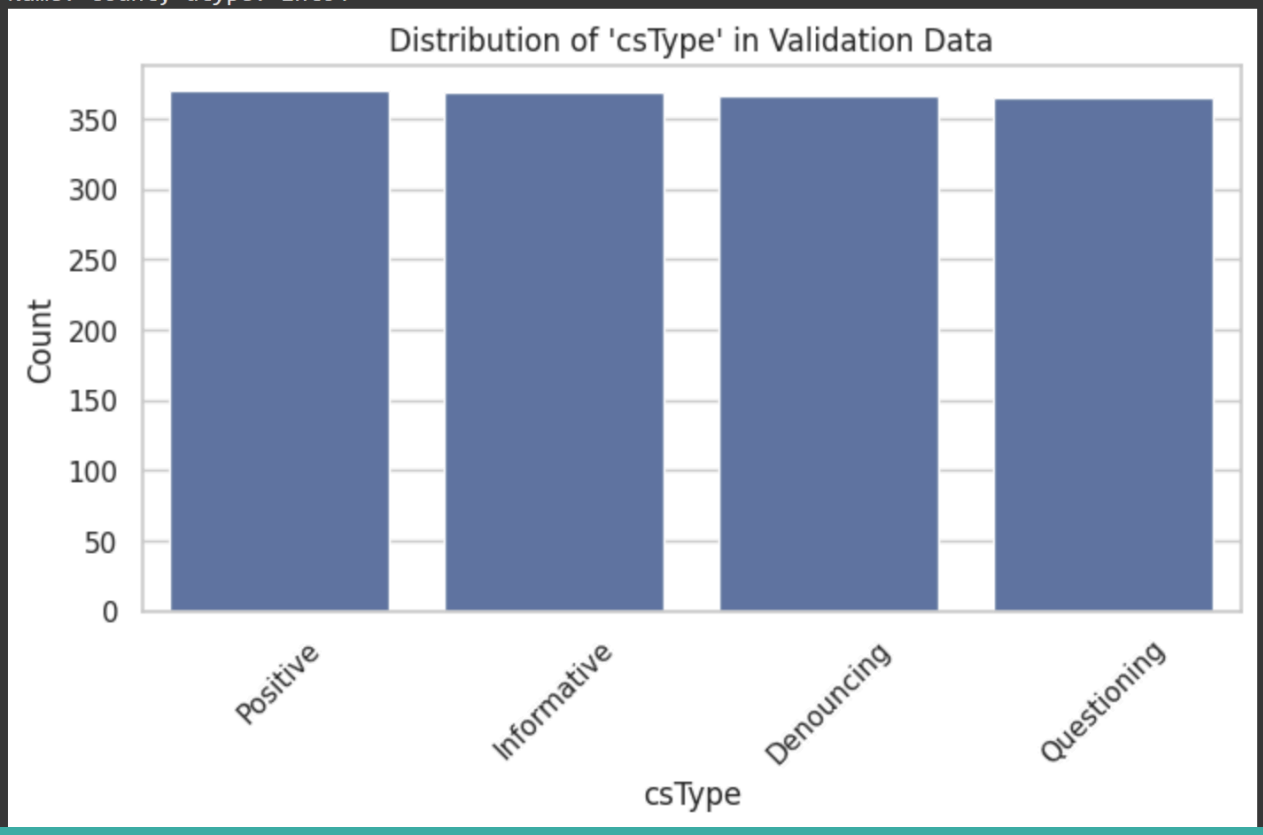
```
=== Validation Data Summary (All Columns) ===
              index                                        hatespeech  \
count    1470.000000                                          1470
unique           NaN                                           900
top              NaN  Migrants rape our young people they must be de...
freq             NaN                                             4
mean     7352.155782                                           NaN
std      4020.244288                                           NaN
min        21.000000                                           NaN
25%      3659.500000                                           NaN
50%      7592.500000                                           NaN
75%     10743.750000                                           NaN
max     13943.000000                                           NaN

          csType                                    counterspeech  \
count       1470                                             1470
unique         4                                             1470
top     Positive  It is essential to promote gender equality and...
freq         370                                                1
mean         NaN                                              NaN
std          NaN                                              NaN
min          NaN                                              NaN
25%          NaN                                              NaN
50%          NaN                                              NaN
75%          NaN                                              NaN
max          NaN                                              NaN
```

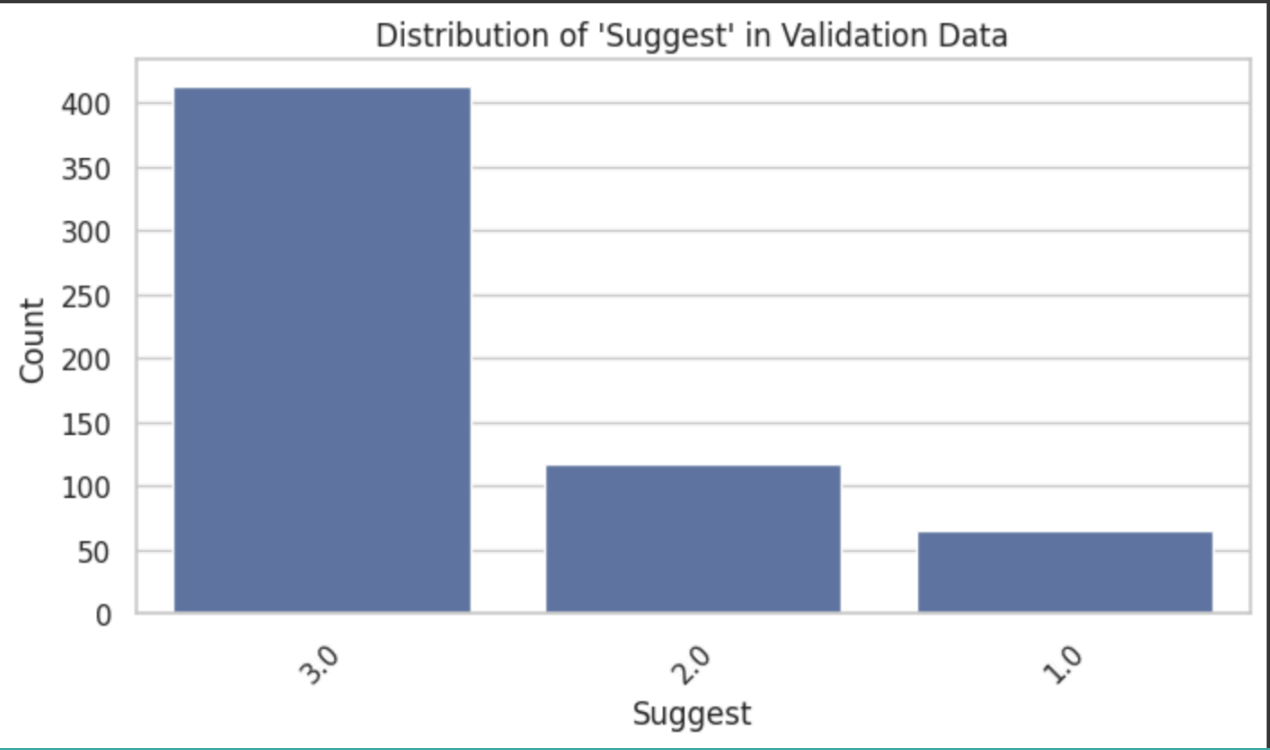
```
            Suggest   Relevance  Aggressive  Complexity  Comments  \
count   596.000000  595.000000  596.000000  596.000000       0.0
unique         NaN         NaN         NaN         NaN       NaN
top            NaN         NaN         NaN         NaN       NaN
freq           NaN         NaN         NaN         NaN       NaN
mean      2.582215    3.379832    2.295302    1.904362       NaN
std       0.682274    1.077906    0.911206    0.660467       NaN
min       1.000000    1.000000    1.000000    1.000000       NaN
25%       2.000000    3.000000    2.000000    2.000000       NaN
50%       3.000000    3.000000    2.000000    2.000000       NaN
75%       3.000000    4.000000    3.000000    2.000000       NaN
max       3.000000    5.000000    5.000000    4.000000       NaN

                source  ... hatespeechTarget powerDynamics  \
count             1470  ...             1470          1470
unique               3  ...               11             8
top     GPT-zeroshot-2  ...          muslims      religion
freq               856  ...              591           792
mean               NaN  ...              NaN           NaN
std                NaN  ...              NaN           NaN
min                NaN  ...              NaN           NaN
25%                NaN  ...              NaN           NaN
50%                NaN  ...              NaN           NaN
75%                NaN  ...              NaN           NaN
max                NaN  ...              NaN           NaN
```

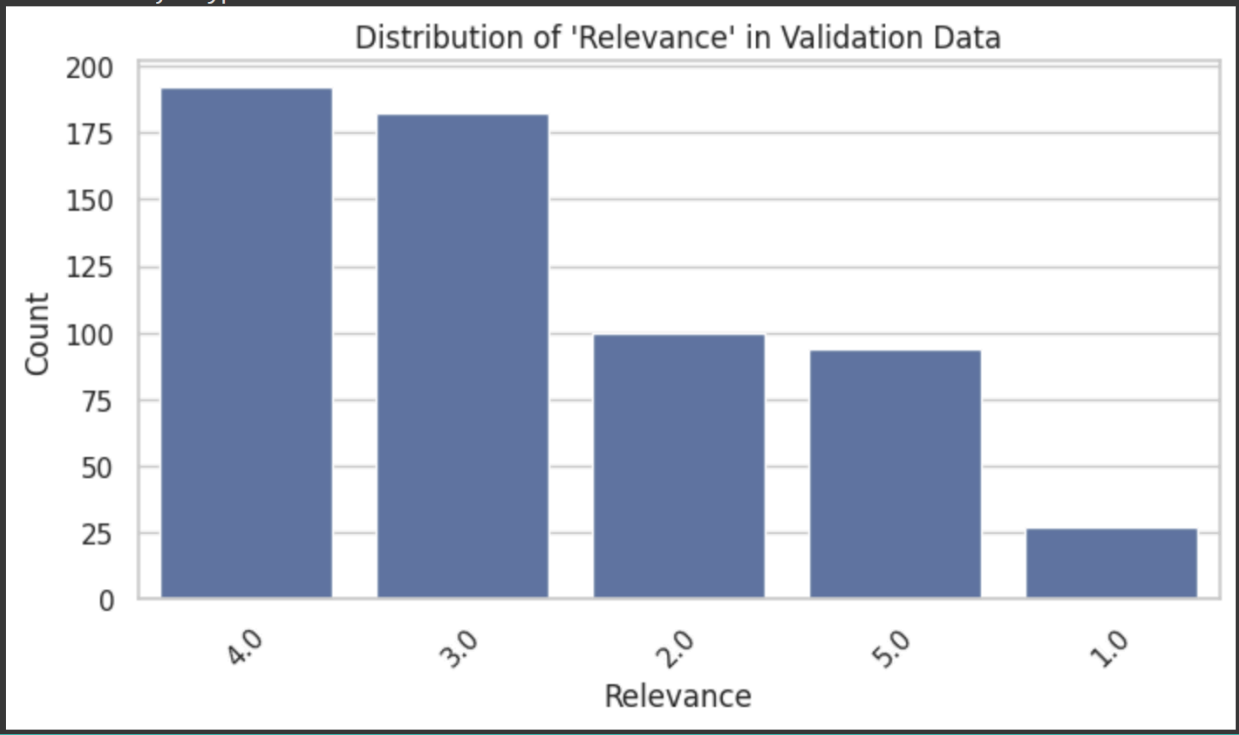
```
Value counts for 'csType' in Validation Data:
csType
Positive       370
Informative    369
Denouncing     366
Questioning    365
Name: count, dtype: int64
```



```
Value counts for 'Suggest' in Validation Data:
Suggest
3.0    413
2.0    117
1.0     66
Name: count, dtype: int64
```



```
Value counts for 'Relevance' in Validation Data:
Relevance
4.0    192
3.0    182
2.0    100
5.0     94
1.0     27
Name: count, dtype: int64
```



```
Column 'Aggresive' not found in Validation Data.

Value counts for 'Complexity' in Validation Data:
Complexity
2.0    373
1.0    147
3.0     62
4.0     14
Name: count, dtype: int64
```