Mini Project Report

on

Hate Speech Data Analytics

Submitted by

Makam Sujith 21BCS061

Ruthvik Jangam 21BCS047

Amballa V. Sriram 21BCS008

Anil Kumar 21BEC056

Under the guidance of

Dr.Animesh Chaturvedi

Designation

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY DHARWAD

29/04/2024

Certificate

This is to certify that the project, entitled **Hate Speech Data Analytics**, is a bonafide record of the Mini Project coursework presented by the students whose names are given below during 2023-24 in partial fulfilment of the requirements of the degree of Bachelor of Technology in Computer Science and Engineering, Electronics and Communication Engineering.

Roll No	Names of Students
<21BCS061>	<Makam Sujith>
<21BCS047>	<Ruthvik Jangam>
<21BCS008>	<Amballa V. Sriram>
<21BEC056>	<Anil Kumar>

Animesh Chaturvedi
(Project Supervisor)

Contents

List of Figures

1 Introduction

Hate speech, defined as any speech, gesture, conduct, writing, or display that may incite violence or prejudicial action against or by a particular individual or group, has become a pervasive issue in the digital age. With the rise of social media platforms and online forums, the dissemination of hate speech has reached unprecedented levels, posing significant challenges to societies worldwide.

In response to this growing concern, researchers and practitioners have increasingly turned to data analytics techniques to understand, identify, and combat hate speech online. By leveraging large datasets extracted from various online platforms, researchers can analyze patterns, trends, and linguistic features associated with hate speech. This analytical approach offers valuable insights into the dynamics of hate speech propagation, its underlying motivations, and its impact on targeted communities.

In this paper, we delve into the realm of hate speech data analytics, exploring the methodologies, challenges, and implications of leveraging data-driven approaches to address this pressing societal issue. We examine the role of language processing, string matching, and computational linguistics in detecting and categorizing hate speech, as well as the ethical considerations and limitations inherent in such endeavors. Furthermore, we discuss the potential applications of hate speech data analytics in informing policy decisions, fostering online safety measures, and promoting inclusive online environments.

Furthermore, ethical considerations are paramount. Striking a balance between protecting free speech and safeguarding individuals from harm is crucial. Automated content moderation systems can inadvertently silence legitimate dissent, while failing to catch all instances of hate speech. Additionally, the potential for data breaches and the misuse of hate speech data for targeted harassment raise significant privacy concerns.

Despite these challenges, hate speech data analytics holds immense potential. Insights gleaned from data analysis can inform policy decisions by providing a deeper understanding of the sources and pervasiveness of hate speech. Additionally, these analyses can be used to develop more effective detection and mitigation tools, fostering safer online environments for all.

By harnessing the power of data and approaching the issue with a nuanced understanding of language and ethics, researchers and practitioners can play a critical role in turning the tide against hate speech. This ongoing battle requires collaboration between researchers, policymakers, technology companies, and civil society to ensure that the digital world becomes a space for inclusive and respectful communication.

2 Related Work

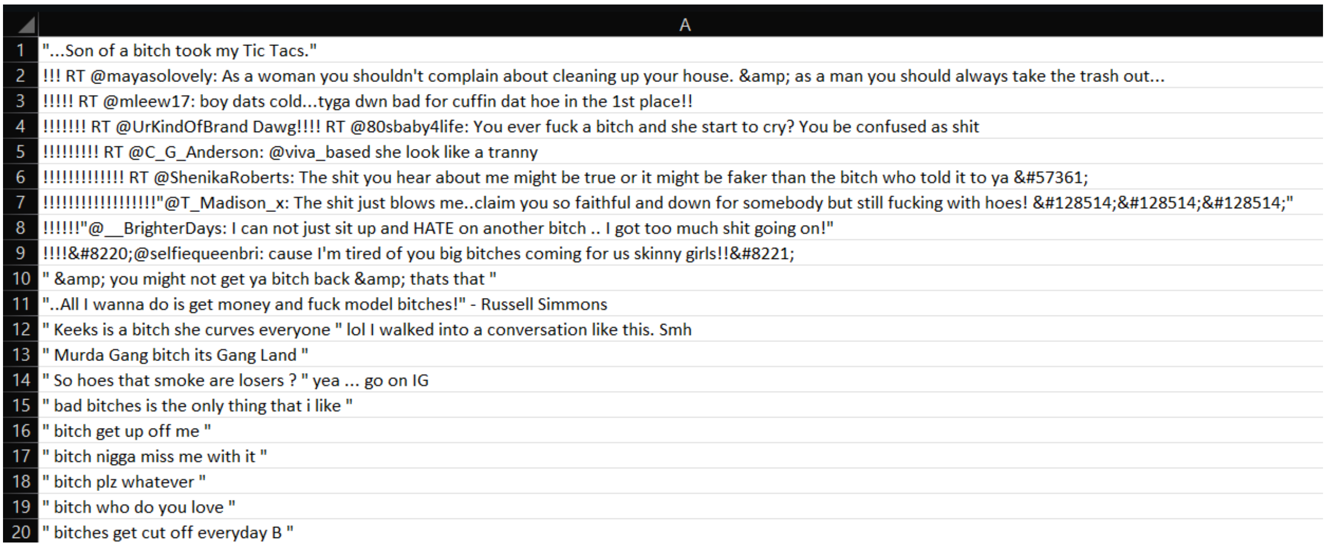
The paper introduces two approaches to quantitatively measure and qualitatively visualize the relationship between co-occurring Hate Terms (HTs). It discusses generating an Inter-agreement HTs list that includes offensive metric values and the HTs lists containing these HTs, aiming to improve hate speech classification efficiency. The paper proposes generating Stable Hate Rules (SHRs) based on frequently co-occurring HTs among multiple hate speech (HS) datasets. It uses metrics like Hatefulness, Relativeness, and Offensiveness to gauge HT severity and presents results using a confusion matrix to show Inter-Agreement between HS data and HTs-lists. Examples of SHRs, concepts, transitivities, and lattices are provided, along with comparisons and analysis of HS data.

Caselli et al., described abuse and offense by studying Inter-Agreement. We have also analyzed Inter-agreement between HTs-list and HS-data Classes, such that, this process generates a Severe HTs-list. Pedersen presented lists of HTs, whereas we have retrieved Severe HTs-list. Liu et al., presented a Multi-Task Deep Neural Network (MT-DNN) for learning text representations for multiple natural language understanding (NLU). Zampieri et al., annotated the datasets for abusive messages named Offensive Language Identification Dataset (OLID), which they further used for Offensive Language in Social Media (Offens Eval), Devlin et al., presented a popular language representation model named Bidirectional Encoder Representations from Transformers (BERT) for tasks like question answering and language inference. al., measured hate and toxicity Recently, Mitos toxicity with the percentage of hate words, and the level of toxicity/inflammatory.

They used HTs available on hatebase.org. which were previously worked out by Hine et al., and Vidgen et al. presented dynamic human and model-based four rounds of hate speech training. Ball et al., studied racial dialect using neural networks' harmful tweet detection. Vidgen et al., described a contextual abuse dataset with six taxonomies for abusive, non-abusive, and neutral speech.

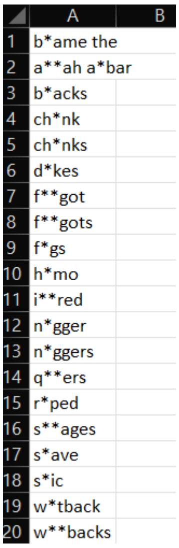
3 Data and Methods

The study collected and analysed two types of data: hate phrase lists and hate speech data. Hate phrase lists were taken from t-davidson's GitHub repository, while hate speech data included offensive remarks sourced from Twitter, notably the Hugging Face dataset. The use of this dataset was approved by the respective user. The hate phrase lists were divided into six categories, while the hate speech data was divided into three sections for study. Using this dataset, analytics were performed to identify patterns and trends in the data. The data appears like this:



	A
1	"...Son of a bitch took my Tic Tacs."
2	!!! RT @mayasolovely: As a woman you shouldn't complain about cleaning up your house. &amp; as a man you should always take the trash out...
3	!!!!! RT @mleew17: boy dats cold...tyga dwn bad for cuffin dat hoe in the 1st place!!
4	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby4life: You ever fuck a bitch and she start to cry? You be confused as shit
5	!!!!!!!!! RT @C_G_Anderson: @viva_based she look like a tranny
6	!!!!!!!!!!!!! RT @ShenikaRoberts: The shit you hear about me might be true or it might be faker than the bitch who told it to ya &#57361;
7	!!!!!!!!!!!!!!!!!!"@T_Madison_x: The shit just blows me..claim you so faithful and down for somebody but still fucking with hoes! &#128514;&#128514;&#128514;"
8	!!!!!!"@__BrighterDays: I can not just sit up and HATE on another bitch .. I got too much shit going on!"
9	!!!!&#8220;@selfiequeenbri: cause I'm tired of you big bitches coming for us skinny girls!!&#8221;
10	" &amp; you might not get ya bitch back &amp; thats that "
11	"..All I wanna do is get money and fuck model bitches!" - Russell Simmons
12	" Keeks is a bitch she curves everyone " lol I walked into a conversation like this. Smh
13	" Murda Gang bitch its Gang Land "
14	" So hoes that smoke are losers ? " yea ... go on IG
15	" bad bitches is the only thing that i like "
16	" bitch get up off me "
17	" bitch nigga miss me with it "
18	" bitch plz whatever "
19	" bitch who do you love "
20	" bitches get cut off everyday B "

Figure 1. dataset containing sentences



	A	B
1	b*ame the	
2	a**ah a*bar	
3	b*acks	
4	ch*nk	
5	ch*nks	
6	d*kes	
7	f**got	
8	f**gots	
9	f*gs	
10	h*mo	
11	i**red	
12	n*gger	
13	n*ggers	
14	q**ers	
15	r*ped	
16	s**ages	
17	s*ave	
18	s*ic	
19	w*tback	
20	w**backs	

Figure 2. List of hate terms

Methods used in the code:

These are the core functions of our project which will contribute more in the project.

MakeListHateTermFrequencies: The makeListHateTermFrequencies method, encapsulated within a class named BasicUtilities, serves the purpose of extracting hate term frequencies from a set of CSV files. These files are named based on different databases and reside within a specified directory. The method iterates through each database file, reads its content, and extracts hate terms along with their frequencies. These terms are then added to a list for further processing.

MakeListHateTermPercentLine: As shown in the above function here we will store the percentage of hate-terms.

IntraHTsListMatrix: This method essentially processes term occurrences, calculates relevant metrics, and writes them to a file, enabling further analysis of hate speech-related data.

InterAgreementTerms: This method calculates and writes inter-agreement terms data to a CSV file based on specified hate term files.

InterAgreementConfusionMatrix: This method calculates various metrics for a confusion matrix, such as counts, percentages, accuracy, recall, precision, and F-measure, based on the provided Hate Term file name (HTFileName). It then writes these metrics to a CSV file representing the confusion matrix for inter-agreement analysis.

SummaryTermConfusionMatrix4Cases: The summaryTermConfusionMatrix4Cases method, declared as private and residing within an undisclosed class, is responsible for generating a summary of confusion matrix statistics for four different cases related to terms. It computes various metrics such as recall and precision for each case and writes the results to a CSV file named ”TermConfusionMatrix.csv”.

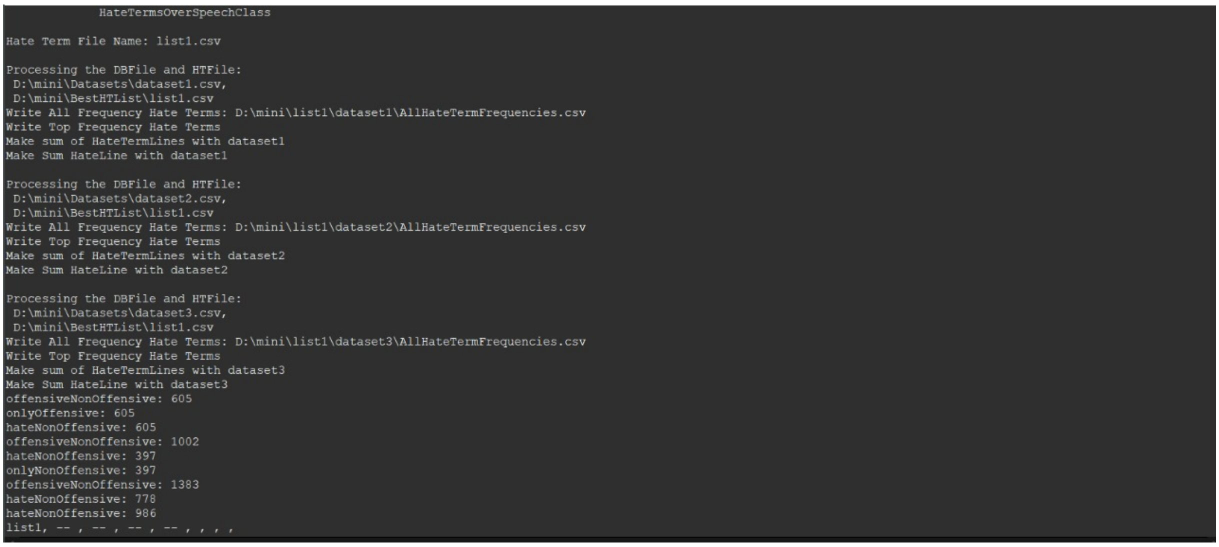
SummaryConfusionMatrix4Cases: The summaryConfusionMatrix4Cases method declared as private and residing within an undisclosed class, is responsible for summarizing confusion matrix statistics for four different cases related to hate, offensive, and non-offensive content classification. It computes various metrics such as recall, precision, accuracy, and F-measure for each case and writes the results to a CSV file named ”ConfusionMatrix.csv”.

MakeOuterJoinHateTermFrequencies: The makeOuterJoinFrequenciesFile method, declared private and residing within a class, is designed to create a string representing the frequencies of a specific term across multiple databases. This method processes CSV files containing hate term frequencies for different databases, extracts the frequency corresponding to the provided term from each file, and constructs a comma-separated string containing these frequencies.

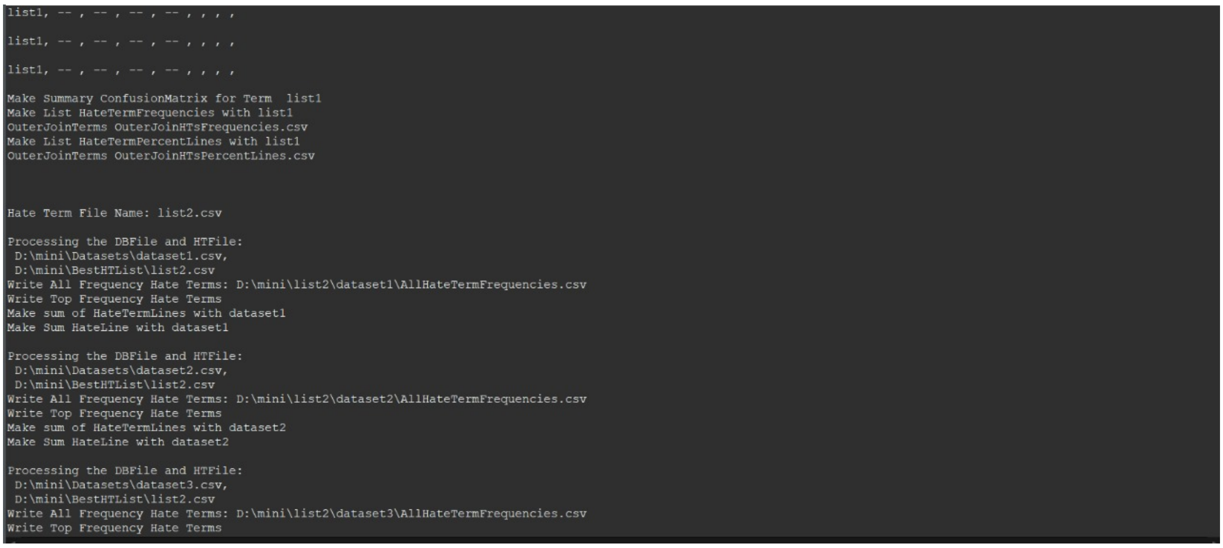
4 Results and Discussions

Output Generated after running the code:

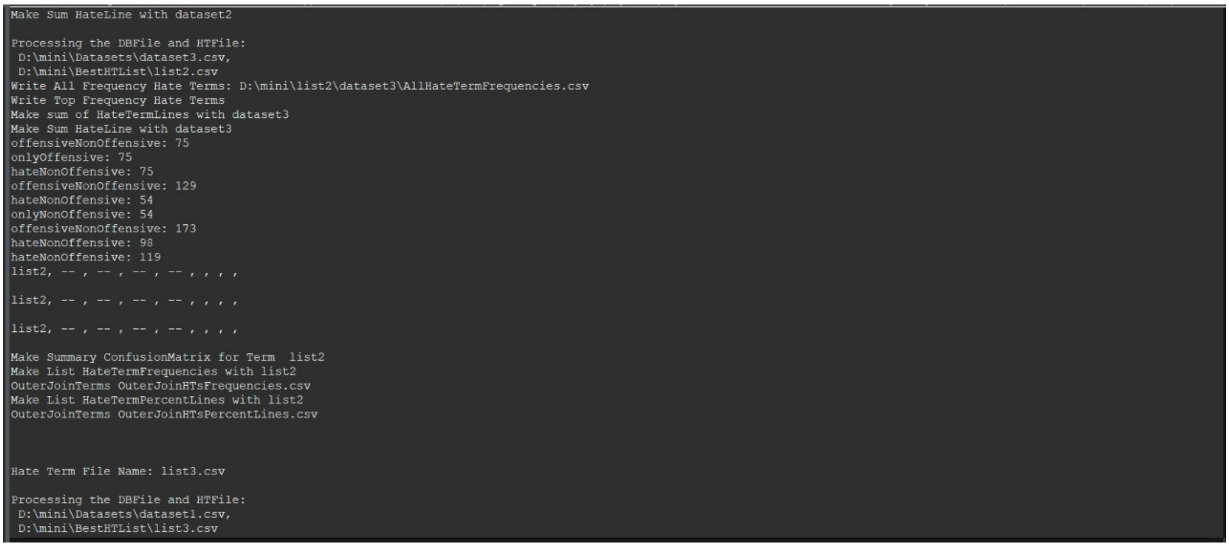
The processing of the hate term file named list1.csv entails iterating through multiple datasets (dataset1.csv, dataset2.csv, dataset3.csv) alongside the same hate term file. For each dataset, the software calculates and records the occurrence frequency of hate terms in corresponding output files. Moreover, it computes aggregate metrics related to hate for each dataset.



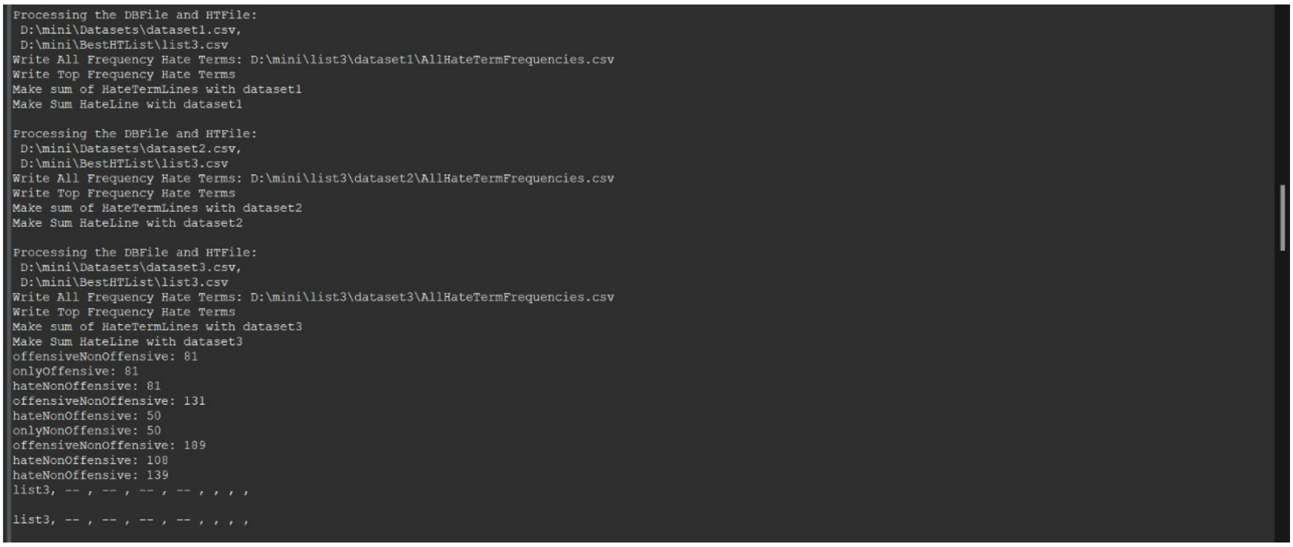
```
            HateTermsOverSpeechClass

Hate Term File Name: list1.csv

Processing the DBFile and HTFile:
 D:\mini\Datasets\dataset1.csv,
 D:\mini\BestHTList\list1.csv
Write All Frequency Hate Terms: D:\mini\list1\dataset1\AllHateTermFrequencies.csv
Write Top Frequency Hate Terms
Make sum of HateTermLines with dataset1
Make Sum HateLine with dataset1

Processing the DBFile and HTFile:
 D:\mini\Datasets\dataset2.csv,
 D:\mini\BestHTList\list1.csv
Write All Frequency Hate Terms: D:\mini\list1\dataset2\AllHateTermFrequencies.csv
Write Top Frequency Hate Terms
Make sum of HateTermLines with dataset2
Make Sum HateLine with dataset2

Processing the DBFile and HTFile:
 D:\mini\Datasets\dataset3.csv,
 D:\mini\BestHTList\list1.csv
Write All Frequency Hate Terms: D:\mini\list1\dataset3\AllHateTermFrequencies.csv
Write Top Frequency Hate Terms
Make sum of HateTermLines with dataset3
Make Sum HateLine with dataset3
offensiveNonOffensive: 605
onlyOffensive: 605
hateNonOffensive: 605
offensiveNonOffensive: 1002
hateNonOffensive: 397
onlyNonOffensive: 397
offensiveNonOffensive: 1383
hateNonOffensive: 778
hateNonOffensive: 986
list1, -- , -- , -- , -- , , , ,
```



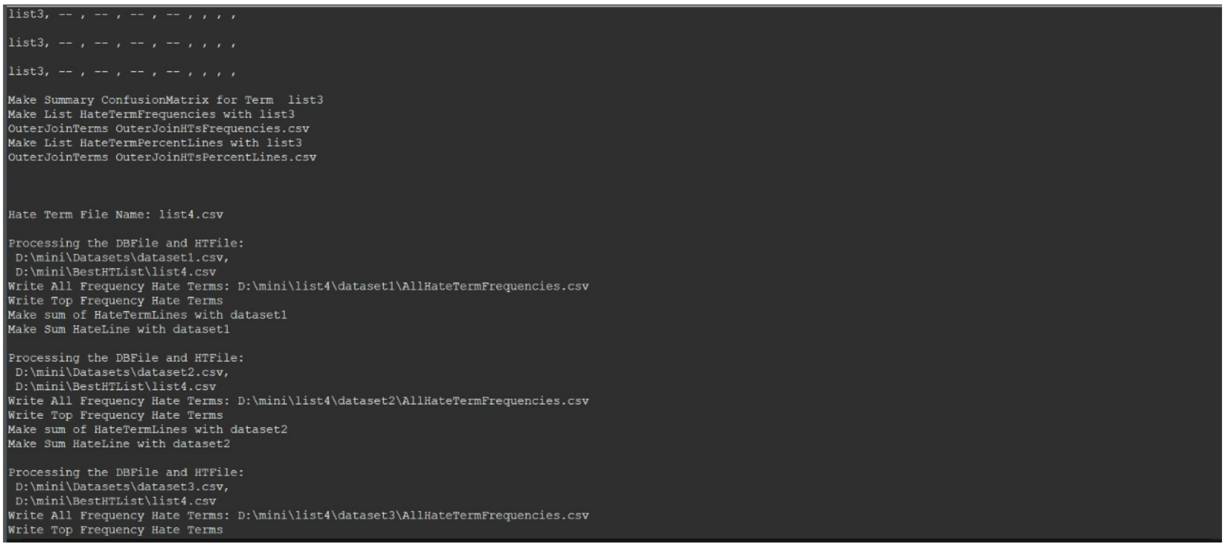
```
list1, -- , -- , -- , -- , , , ,

list1, -- , -- , -- , -- , , , ,

list1, -- , -- , -- , -- , , , ,

Make Summary ConfusionMatrix for Term  list1
Make List HateTermFrequencies with list1
OuterJoinTerms OuterJoinHTsFrequencies.csv
Make List HateTermPercentLines with list1
OuterJoinTerms OuterJoinHTsPercentLines.csv



Hate Term File Name: list2.csv

Processing the DBFile and HTFile:
 D:\mini\Datasets\dataset1.csv,
 D:\mini\BestHTList\list2.csv
Write All Frequency Hate Terms: D:\mini\list2\dataset1\AllHateTermFrequencies.csv
Write Top Frequency Hate Terms
Make sum of HateTermLines with dataset1
Make Sum HateLine with dataset1

Processing the DBFile and HTFile:
 D:\mini\Datasets\dataset2.csv,
 D:\mini\BestHTList\list2.csv
Write All Frequency Hate Terms: D:\mini\list2\dataset2\AllHateTermFrequencies.csv
Write Top Frequency Hate Terms
Make sum of HateTermLines with dataset2
Make Sum HateLine with dataset2

Processing the DBFile and HTFile:
 D:\mini\Datasets\dataset3.csv,
 D:\mini\BestHTList\list2.csv
Write All Frequency Hate Terms: D:\mini\list2\dataset3\AllHateTermFrequencies.csv
Write Top Frequency Hate Terms
```



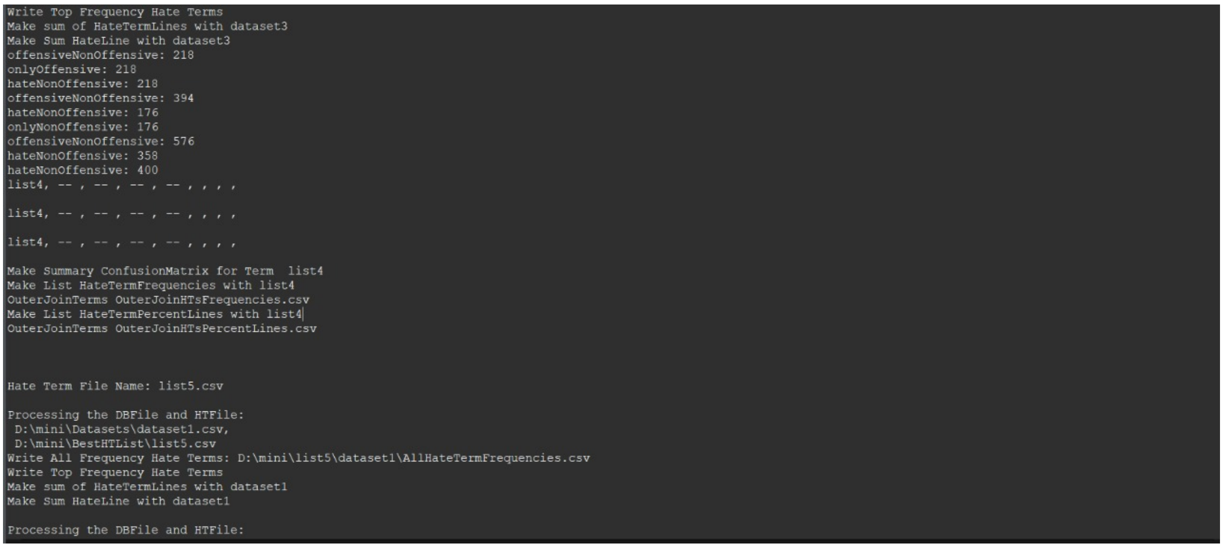
```
Make Sum HateLine with dataset2

Processing the DBFile and HTFile:
 D:\mini\Datasets\dataset3.csv,
 D:\mini\BestHTList\list2.csv
Write All Frequency Hate Terms: D:\mini\list2\dataset3\AllHateTermFrequencies.csv
Write Top Frequency Hate Terms
Make sum of HateTermLines with dataset3
Make Sum HateLine with dataset3
offensiveNonOffensive: 75
onlyOffensive: 75
hateNonOffensive: 75
offensiveNonOffensive: 129
hateNonOffensive: 54
onlyNonOffensive: 54
offensiveNonOffensive: 173
hateNonOffensive: 98
hateNonOffensive: 119
list2, -- , -- , -- , -- , , , ,

list2, -- , -- , -- , -- , , , ,

list2, -- , -- , -- , -- , , , ,

Make Summary ConfusionMatrix for Term  list2
Make List HateTermFrequencies with list2
OuterJoinTerms OuterJoinHTsFrequencies.csv
Make List HateTermPercentLines with list2
OuterJoinTerms OuterJoinHTsPercentLines.csv



Hate Term File Name: list3.csv

Processing the DBFile and HTFile:
 D:\mini\Datasets\dataset1.csv,
 D:\mini\BestHTList\list3.csv
```



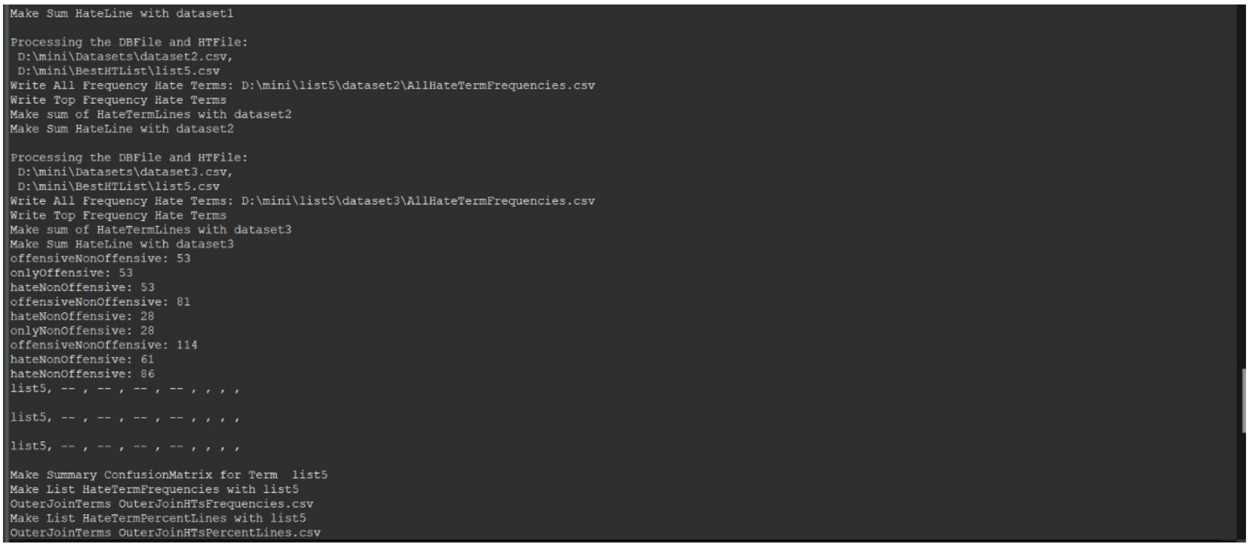
```
Processing the DBFile and HTFile:
 D:\mini\Datasets\dataset1.csv,
 D:\mini\BestHTList\list3.csv
Write All Frequency Hate Terms: D:\mini\list3\dataset1\AllHateTermFrequencies.csv
Write Top Frequency Hate Terms
Make sum of HateTermLines with dataset1
Make Sum HateLine with dataset1

Processing the DBFile and HTFile:
 D:\mini\Datasets\dataset2.csv,
 D:\mini\BestHTList\list3.csv
Write All Frequency Hate Terms: D:\mini\list3\dataset2\AllHateTermFrequencies.csv
Write Top Frequency Hate Terms
Make sum of HateTermLines with dataset2
Make Sum HateLine with dataset2

Processing the DBFile and HTFile:
 D:\mini\Datasets\dataset3.csv,
 D:\mini\BestHTList\list3.csv
Write All Frequency Hate Terms: D:\mini\list3\dataset3\AllHateTermFrequencies.csv
Write Top Frequency Hate Terms
Make sum of HateTermLines with dataset3
Make Sum HateLine with dataset3
offensiveNonOffensive: 81
onlyOffensive: 81
hateNonOffensive: 81
offensiveNonOffensive: 131
hateNonOffensive: 50
onlyNonOffensive: 50
offensiveNonOffensive: 189
hateNonOffensive: 108
hateNonOffensive: 139
list3, -- , -- , -- , -- , , , ,

list3, -- , -- , -- , -- , , , ,
```



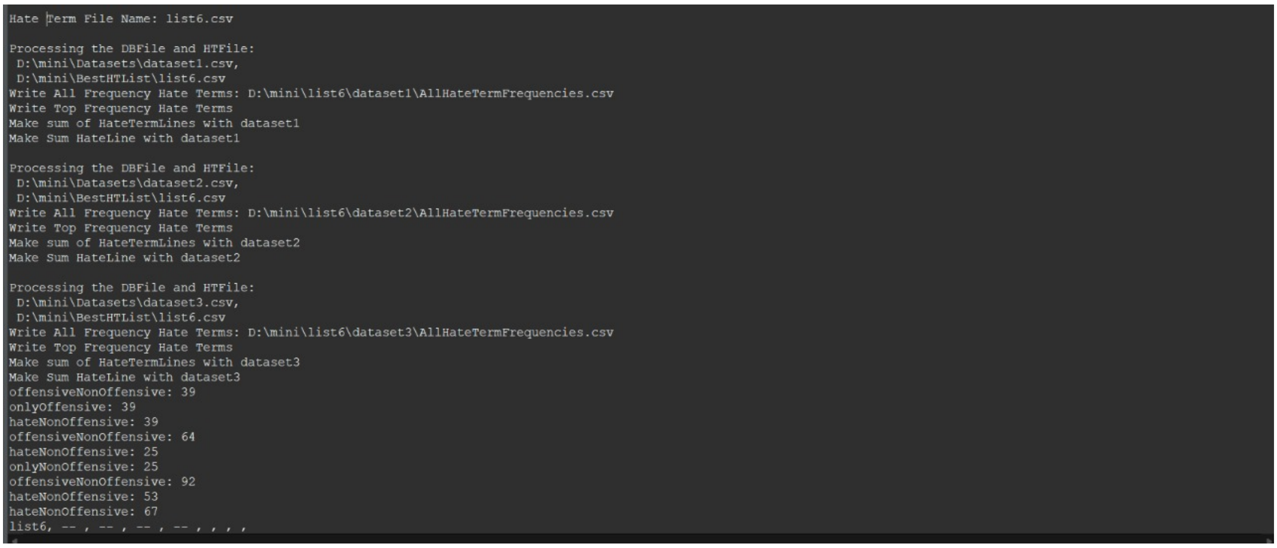
```
list3, -- , -- , -- , -- , , , ,

list3, -- , -- , -- , -- , , , ,

list3, -- , -- , -- , -- , , , ,

Make Summary ConfusionMatrix for Term  list3
Make List HateTermFrequencies with list3
OuterJoinTerms OuterJoinHTsFrequencies.csv
Make List HateTermPercentLines with list3
OuterJoinTerms OuterJoinHTsPercentLines.csv



Hate Term File Name: list4.csv

Processing the DBFile and HTFile:
 D:\mini\Datasets\dataset1.csv,
 D:\mini\BestHTList\list4.csv
Write All Frequency Hate Terms: D:\mini\list4\dataset1\AllHateTermFrequencies.csv
Write Top Frequency Hate Terms
Make sum of HateTermLines with dataset1
Make Sum HateLine with dataset1

Processing the DBFile and HTFile:
 D:\mini\Datasets\dataset2.csv,
 D:\mini\BestHTList\list4.csv
Write All Frequency Hate Terms: D:\mini\list4\dataset2\AllHateTermFrequencies.csv
Write Top Frequency Hate Terms
Make sum of HateTermLines with dataset2
Make Sum HateLine with dataset2

Processing the DBFile and HTFile:
 D:\mini\Datasets\dataset3.csv,
 D:\mini\BestHTList\list4.csv
Write All Frequency Hate Terms: D:\mini\list4\dataset3\AllHateTermFrequencies.csv
Write Top Frequency Hate Terms
```



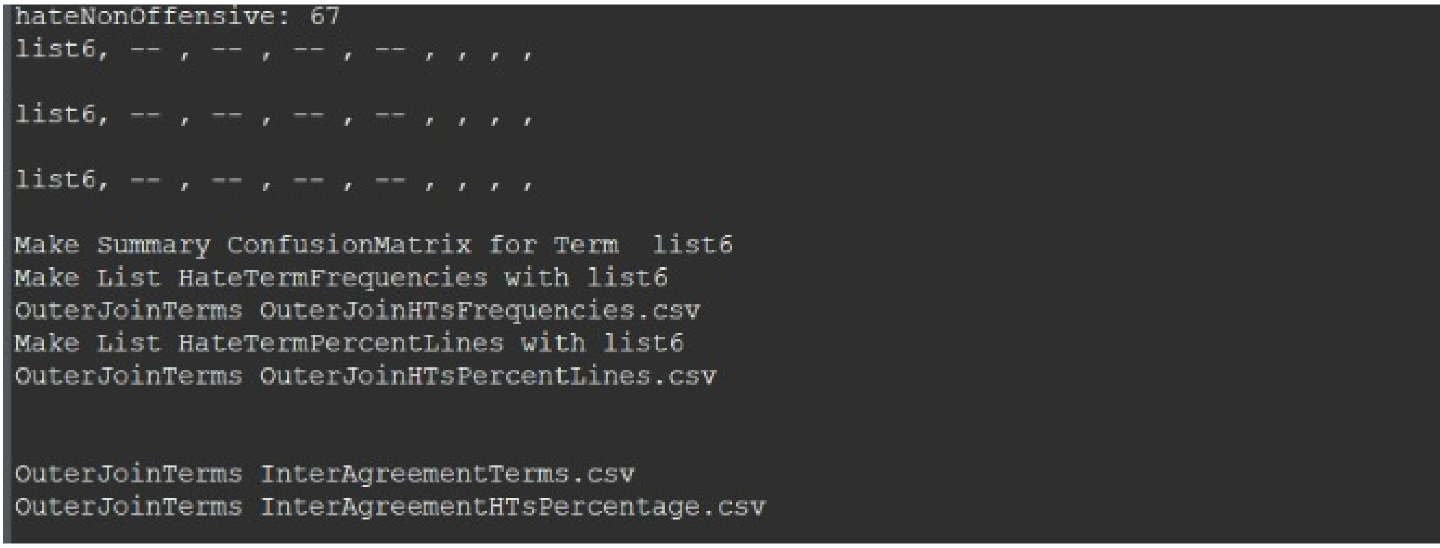
```
Write Top Frequency Hate Terms
Make sum of HateTermLines with dataset3
Make Sum HateLine with dataset3
offensiveNonOffensive: 218
onlyOffensive: 218
hateNonOffensive: 218
offensiveNonOffensive: 394
hateNonOffensive: 176
onlyNonOffensive: 176
offensiveNonOffensive: 576
hateNonOffensive: 358
hateNonOffensive: 400
list4, -- , -- , -- , -- , , , ,

list4, -- , -- , -- , -- , , , ,

list4, -- , -- , -- , -- , , , ,

Make Summary ConfusionMatrix for Term  list4
Make List HateTermFrequencies with list4
OuterJoinTerms OuterJoinHTsFrequencies.csv
Make List HateTermPercentLines with list4
OuterJoinTerms OuterJoinHTsPercentLines.csv



Hate Term File Name: list5.csv

Processing the DBFile and HTFile:
 D:\mini\Datasets\dataset1.csv,
 D:\mini\BestHTList\list5.csv
Write All Frequency Hate Terms: D:\mini\list5\dataset1\AllHateTermFrequencies.csv
Write Top Frequency Hate Terms
Make sum of HateTermLines with dataset1
Make Sum HateLine with dataset1

Processing the DBFile and HTFile:
```



```
Make Sum HateLine with dataset1

Processing the DBFile and HTFile:
 D:\mini\Datasets\dataset2.csv,
 D:\mini\BestHTList\list5.csv
Write All Frequency Hate Terms: D:\mini\list5\dataset2\AllHateTermFrequencies.csv
Write Top Frequency Hate Terms
Make sum of HateTermLines with dataset2
Make Sum HateLine with dataset2

Processing the DBFile and HTFile:
 D:\mini\Datasets\dataset3.csv,
 D:\mini\BestHTList\list5.csv
Write All Frequency Hate Terms: D:\mini\list5\dataset3\AllHateTermFrequencies.csv
Write Top Frequency Hate Terms
Make sum of HateTermLines with dataset3
Make Sum HateLine with dataset3
offensiveNonOffensive: 53
onlyOffensive: 53
hateNonOffensive: 53
offensiveNonOffensive: 81
hateNonOffensive: 28
onlyNonOffensive: 28
offensiveNonOffensive: 114
hateNonOffensive: 61
hateNonOffensive: 86
list5, -- , -- , -- , -- , , , ,

list5, -- , -- , -- , -- , , , ,

list5, -- , -- , -- , -- , , , ,

Make Summary ConfusionMatrix for Term  list5
Make List HateTermFrequencies with list5
OuterJoinTerms OuterJoinHTsFrequencies.csv
Make List HateTermPercentLines with list5
OuterJoinTerms OuterJoinHTsPercentLines.csv
```



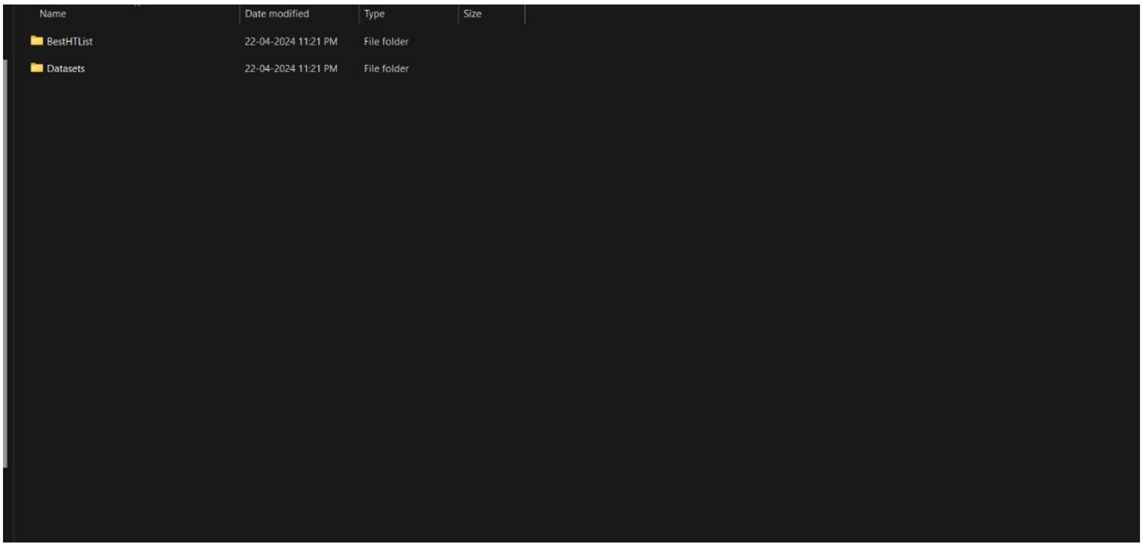
```
Hate Term File Name: list6.csv

Processing the DBFile and HTFile:
 D:\mini\Datasets\dataset1.csv,
 D:\mini\BestHTList\list6.csv
Write All Frequency Hate Terms: D:\mini\list6\dataset1\AllHateTermFrequencies.csv
Write Top Frequency Hate Terms
Make sum of HateTermLines with dataset1
Make Sum HateLine with dataset1

Processing the DBFile and HTFile:
 D:\mini\Datasets\dataset2.csv,
 D:\mini\BestHTList\list6.csv
Write All Frequency Hate Terms: D:\mini\list6\dataset2\AllHateTermFrequencies.csv
Write Top Frequency Hate Terms
Make sum of HateTermLines with dataset2
Make Sum HateLine with dataset2

Processing the DBFile and HTFile:
 D:\mini\Datasets\dataset3.csv,
 D:\mini\BestHTList\list6.csv
Write All Frequency Hate Terms: D:\mini\list6\dataset3\AllHateTermFrequencies.csv
Write Top Frequency Hate Terms
Make sum of HateTermLines with dataset3
Make Sum HateLine with dataset3
offensiveNonOffensive: 39
onlyOffensive: 39
hateNonOffensive: 39
offensiveNonOffensive: 64
hateNonOffensive: 25
onlyNonOffensive: 25
offensiveNonOffensive: 92
hateNonOffensive: 53
hateNonOffensive: 67
list6, -- , -- , -- , -- , , , ,
```



```
hateNonOffensive: 67
list6, -- , -- , -- , -- , , , ,

list6, -- , -- , -- , -- , , , ,

list6, -- , -- , -- , -- , , , ,

Make Summary ConfusionMatrix for Term  list6
Make List HateTermFrequencies with list6
OuterJoinTerms OuterJoinHTsFrequencies.csv
Make List HateTermPercentLines with list6
OuterJoinTerms OuterJoinHTsPercentLines.csv


OuterJoinTerms InterAgreementTerms.csv
OuterJoinTerms InterAgreementHTsPercentage.csv
```

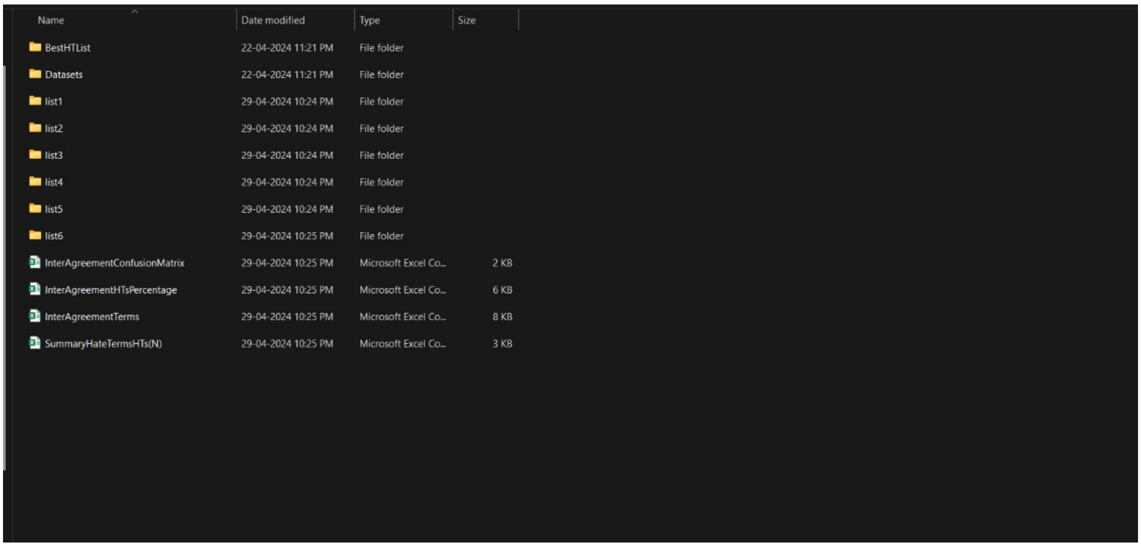
Once all datasets are processed, the program generates a summary confusion matrix and various lists based on hate term frequencies and percentages specific to the term list1.

From here we will show the folder structure of our code before and after execution.



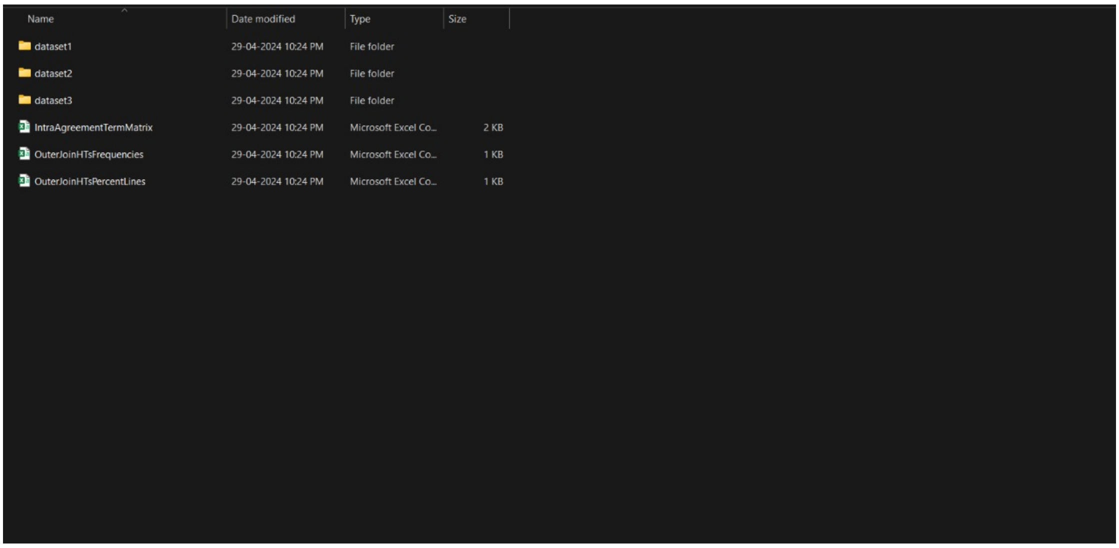
Name	Date modified	Type	Size
BestHTList	22-04-2024 11:21 PM	File folder	
Datasets	22-04-2024 11:21 PM	File folder	

Figure 3. folders before code execution



Name	Date modified	Type	Size
BestHTList	22-04-2024 11:21 PM	File folder	
Datasets	22-04-2024 11:21 PM	File folder	
list1	29-04-2024 10:24 PM	File folder	
list2	29-04-2024 10:24 PM	File folder	
list3	29-04-2024 10:24 PM	File folder	
list4	29-04-2024 10:24 PM	File folder	
list5	29-04-2024 10:24 PM	File folder	
list6	29-04-2024 10:25 PM	File folder	
InterAgreementConfusionMatrix	29-04-2024 10:25 PM	Microsoft Excel Co...	2 KB
InterAgreementHTsPercentage	29-04-2024 10:25 PM	Microsoft Excel Co...	6 KB
InterAgreementTerms	29-04-2024 10:25 PM	Microsoft Excel Co...	8 KB
SummaryHateTermsHTs(N)	29-04-2024 10:25 PM	Microsoft Excel Co...	3 KB

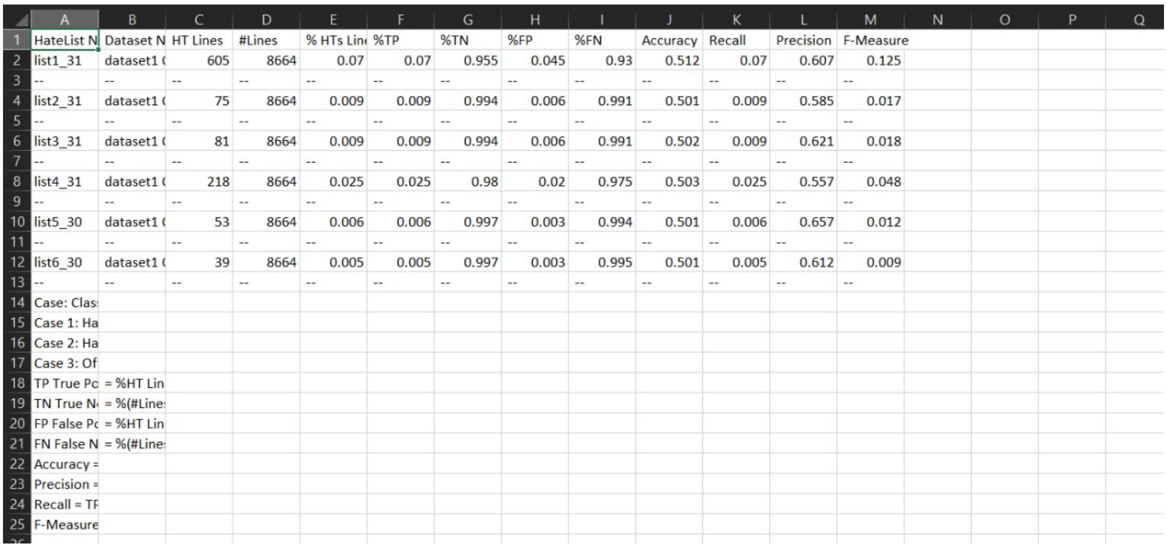
Figure 4. folders after execution



Name	Date modified	Type	Size
dataset1	29-04-2024 10:24 PM	File folder	
dataset2	29-04-2024 10:24 PM	File folder	
dataset3	29-04-2024 10:24 PM	File folder	
IntraAgreementTermMatrix	29-04-2024 10:24 PM	Microsoft Excel Co...	2 KB
OuterJoinHTsFrequencies	29-04-2024 10:24 PM	Microsoft Excel Co...	1 KB
OuterJoinHTsPercentLines	29-04-2024 10:24 PM	Microsoft Excel Co...	1 KB

Figure 5. Files generated for each list

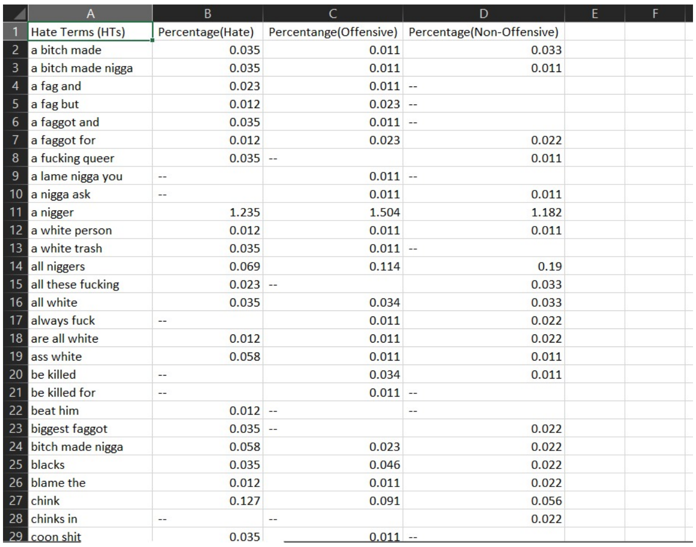
The output that we are focusing on the most is the **inter-agreement confusion matrix**. This provides details on the confusion matrix elements including True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) used to calculate accuracy, precision, recall, and f-measure for each hate term in the entire hate terms list within hate speech data.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	HateList N	Dataset N	HT Lines	#Lines	% HTs Lin	%TP	%TN	%FP	%FN	Accuracy	Recall	Precision	F-Measure
2	list1_31	dataset1 (	605	8664	0.07	0.07	0.955	0.045	0.93	0.512	0.07	0.607	0.125
3	--	--	--	--	--	--	--	--	--	--	--	--	--
4	list2_31	dataset1 (	75	8664	0.009	0.009	0.994	0.006	0.991	0.501	0.009	0.585	0.017
5	--	--	--	--	--	--	--	--	--	--	--	--	--
6	list3_31	dataset1 (	81	8664	0.009	0.009	0.994	0.006	0.991	0.502	0.009	0.621	0.018
7	--	--	--	--	--	--	--	--	--	--	--	--	--
8	list4_31	dataset1 (	218	8664	0.025	0.025	0.98	0.02	0.975	0.503	0.025	0.557	0.048
9	--	--	--	--	--	--	--	--	--	--	--	--	--
10	list5_30	dataset1 (	53	8664	0.006	0.006	0.997	0.003	0.994	0.501	0.006	0.657	0.012
11	--	--	--	--	--	--	--	--	--	--	--	--	--
12	list6_30	dataset1 (	39	8664	0.005	0.005	0.997	0.003	0.995	0.501	0.005	0.612	0.009
13	--	--	--	--	--	--	--	--	--	--	--	--	--
14	Case: Clas												
15	Case 1: Ha												
16	Case 2: Ha												
17	Case 3: Of												
18	TP True Po	= %HT Lin											
19	TN True N	= %(#Line											
20	FP False Po	= %HT Lin											
21	FN False N	= %(#Line											
22	Accuracy =												
23	Precision =												
24	Recall = TF												
25	F-Measure												

Figure 6. inter-agreement confusion matrix

For our hate speech dataset, we merged hate terms from various lists to establish consensus. We then evaluated each term's offensiveness based on its level of hatred and relevance. Some terms were found to be less offensive when used alone. We calculated corresponding percentages to reflect these findings.



	A	B	C	D
1	Hate Terms (HTs)	Percentage(Hate)	Percentange(Offensive)	Percentage(Non-Offensive)
2	a bitch made	0.035	0.011	0.033
3	a bitch made nigga	0.035	0.011	0.011
4	a fag and	0.023	0.011	--
5	a fag but	0.012	0.023	--
6	a faggot and	0.035	0.011	--
7	a faggot for	0.012	0.023	0.022
8	a fucking queer	0.035	--	0.011
9	a lame nigga you	--	0.011	--
10	a nigga ask	--	0.011	0.011
11	a nigger	1.235	1.504	1.182
12	a white person	0.012	0.011	0.011
13	a white trash	0.035	0.011	--
14	all niggers	0.069	0.114	0.19
15	all these fucking	0.023	--	0.033
16	all white	0.035	0.034	0.033
17	always fuck	--	0.011	0.022
18	are all white	0.012	0.011	0.022
19	ass white	0.058	0.011	0.011
20	be killed	--	0.034	0.011
21	be killed for	--	0.011	--
22	beat him	0.012	--	--
23	biggest faggot	0.035	--	0.022
24	bitch made nigga	0.058	0.023	0.022
25	blacks	0.035	0.046	0.022
26	blame the	0.012	0.011	0.022
27	chink	0.127	0.091	0.056
28	chinks in	--	--	0.022
29	coon shit	0.035	0.011	--

Figure 7. Inter agreement Hate Term matrix

5 Conclusion

Based on the comprehensive analysis conducted on hate speech data, several key findings have emerged. Through the examination of inter-agreement between hate terms (HTs) and hate speech data (HS-data), valuable insights into the frequency, offensiveness, and patterns of hate speech have been generated. The identification of Stable Hate Rules (SHRs) and associated concepts has allowed for the establishment of transitivity relationships among hate terms, facilitating a deeper understanding of context-specific hate speech dynamics.

Quantitative analysis, employing a proposed threshold for severe hate terms, has demonstrated the superior performance of the Severe HTs-list in capturing and categorizing hate speech. Meanwhile, qualitative analysis utilizing SHRs has provided visual representations of hate term co-occurrences, offering nuanced insights into the underlying dynamics of hate speech propagation. for all users.

These findings not only contribute to the body of knowledge on hate speech analytics but also offer practical implications for the development of more effective moderation strategies and intervention measures. By leveraging the insights gleaned from this analysis, stakeholders can better address the challenges posed by online hate speech, fostering a safer and more inclusive digital environment for all users.

quantitative analysis using a newly proposed threshold for classifying ”severe” hate terms has demonstrated the effectiveness of a refined ”Severe HTs-list” in capturing and categorizing the most egregious forms of hate speech. This refined list can be a powerful tool for identifying and filtering out the most harmful content.

By understanding the structure and patterns of hate speech, stakeholders like social media platforms and policymakers can create more targeted and efficient methods to combat online hate.

Ultimately, these insights pave the way for a safer and more inclusive digital environment. By leveraging the knowledge gleaned from this data analysis, we can work towards a future where online spaces foster respectful and constructive communication for all users.

References

1. A. Chaturvedi, A. Tiwari and N. Spyratos, ”minStab: Stable Network Evolution Rule Mining for System Changeability Analysis,” in IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 5, no. 2, pp. 274-283, April 2021, doi: 10.1109/TETCI.2019.2892734.

2. A. Chaturvedi and R. Sharma, ”minOffense: Inter-Agreement Hate Terms for Stable Rules, Concepts, Transitivities, and Lattices,” 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA), Shenzhen, China, 2022, pp. 1-10, doi: 10.1109/DSAA54385.2022.10032389.