The 9th IEEE International Conference on Data Science and Advanced Analytics

minOffense: Inter-Agreement Hate Terms for Stable Rules, Concepts, Transitivities, and Lattices

Dr. Animesh Chaturvedi
Indian Institute of Information Technology Dharwad,
Dharwad, Karnataka, India
animesh.chaturvedi88@gmail.com

Dr. Rajesh Sharma
University of Tartu,
Tartu, Estonia
rajesh.sharma@ut.ee

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY

IEEE | Cosponsers ACM/ASA/CCF
DSAA 2022
http://dsaa2022.dsaa.co

Motivation and Research Questions

Motivation and Problem

- For a given set of Hate Terms lists (HTs-lists) and Hate Speech data (HS-data), it is challenging to understand which hate term contributes the most for hate speech.
- Two approaches to the relationship between co-occurring Hate Terms (HTs).

1. Quantitative Analysis

- To create an ***Inter-agreement HTs-list***, which explains the contribution of an individual hate term toward hate speech.
- To produce a **Severe Hate Terms list** (***Severe HTs-list***)

2. Qualitatively Analysis

- ***Stable Hate Rule*** (***SHR***) mining detects ordered frequently co-occurring HTs with ***minimum Stability*** (***minStab***). This form ***Stable Hate Rules*** and ***Concepts***.
- These rules and concepts are used to visualise the graphs of ***Transitivities*** and ***Lattices*** formed by HTs.

Research Questions

- **RQ1**: How to perform ***Inter-agreement analysis***, which provide information about common HTs between a HS-data and multiple HTs-lists?
- **RQ2**: How to use an Inter-agreement HTs-list to generate a ***Severe HTs-list*** for efficient Hate Speech classification?
- **RQ3:** How much better classification is achieved using the Severe HTs-list compared to any of the given HTs-lists?
- **RQ4a:** How to generate ***Stable Hate Rules*** (**SHRs**) that represent frequently co-occurring HTs among multiple HS-data?
- **RQ4b:** How to make hate concepts and visualise the relationship between co-occurring HTs from SHRs?

Quantitative analysis: Inter-Agreement and Severe Hate Terms lists

1. **Quantitative Analysis**

- **To create an *Inter-agreement HTs-list*, which explains the contribution of an individual hate term toward hate speech.**
- **To produce a Severe Hate Terms list (*Severe HTs-list*)**

2. Qualitatively Analysis

- *Stable Hate Rule* (*SHR*) mining detects ordered frequently co-occurring HTs with *minimum Stability* (*minStab*). This form *Stable Hate Rules* and *Concepts*.
- These rules and concepts are used to visualise the graphs of *Transitivities* and *Lattices* formed by HTs.

Overview

- Inspired by the concepts of Shapley value
  - the contribution by individual players in a game
  - the contribution of an individual HT towards hate speech
- Three classes of Hate Speech
  - **Hate:** class indicates the lines definitely contain HTs.
  - **Relative-hate:** class indicates the lines contain mild HTs.
  - **No-hate:** class indicates the lines do not contain HTs.
- Proposed metrics: **Hatefulness**, **Relativeness**, and **Offensiveness**
- To make ***Inter-agreement HTs-list***
- To measure the severity of HTs and generate ***Severe Hate Terms list***

Single Hate Terms List Analysis

4 Artifacts

1. Creation of hate terms frequencies
2. AllHateTermsFrequencies and TopTermsFrequency
3. AllHTsPercentLine
4. OuterJoinHTsFrequencies and OuterJoinHTsPercentLines

Discussed with experiments

Intra-Agreement-HTs for each HTs-list (5^{th} Artifact)

- **Intra-Agreement between a HTs-list and a HS-data**

$$\text{Hatefulness} = \{1 \text{ or } 0 \mid \text{HT} \in \text{Hate class or not, respectively}\}$$

$$\text{Relativeness (Hate)} = \frac{\text{FreqHT in Hate Class}}{\text{FreqHT in Relative-hate class and FreqHT in No-hate class}}$$

$$\text{Relativeness (Hate + Relative-hate)} = \frac{\text{FreqHT in Hate Class} + \text{FreqHT in Relative-hate class}}{\text{FreqHT in No-hate class}}$$

Useful for Inter-Agreement analysis of Multiple HTs-list

Multiple Hate Terms Lists analysis

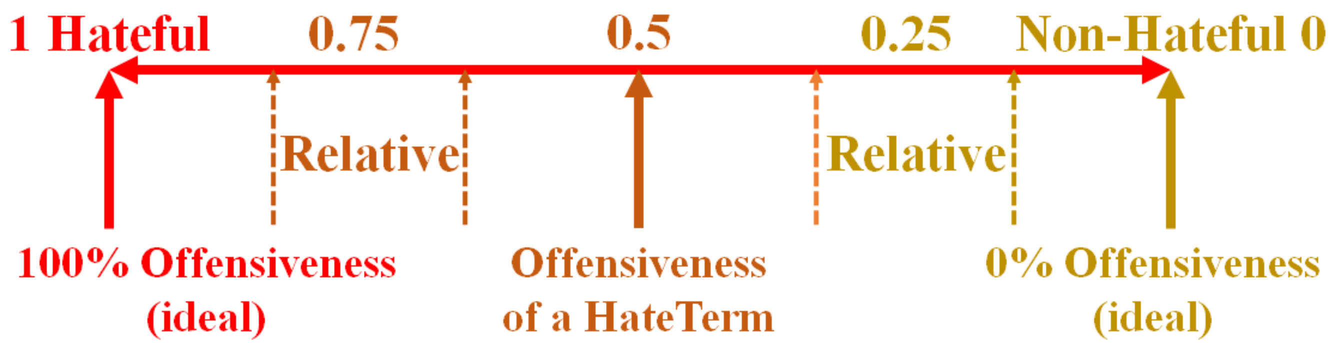
Inter-agreement Hate Terms Analysis (Answer to RQ1)

- Agreement between
  - the HS-data and
  - the multiple **HTs-lists = {HTs-list1, HTs-list2,... HTs-listN}**
- Inter-Agreement HTs Analysis as a matrix **IA** of size **N** × **M**,
  - **N** represents the number of HTs-lists and **M** number of classes in a HS-data.
  - $\mathbf{IA}_{\mathbf{ij}}$ represents the information about HTs of a given HTs-lists, which are present in a class of HS-data.
- Generate a ***Inter-agreement HTs-list*** containing HTs with two kinds of information
  - It contains Offensiveness metric value of each HT in the HS-data.
  - It mentions the HTs-lists which contains those HTs.

Inter-Agreement HTs (6th Artifact)

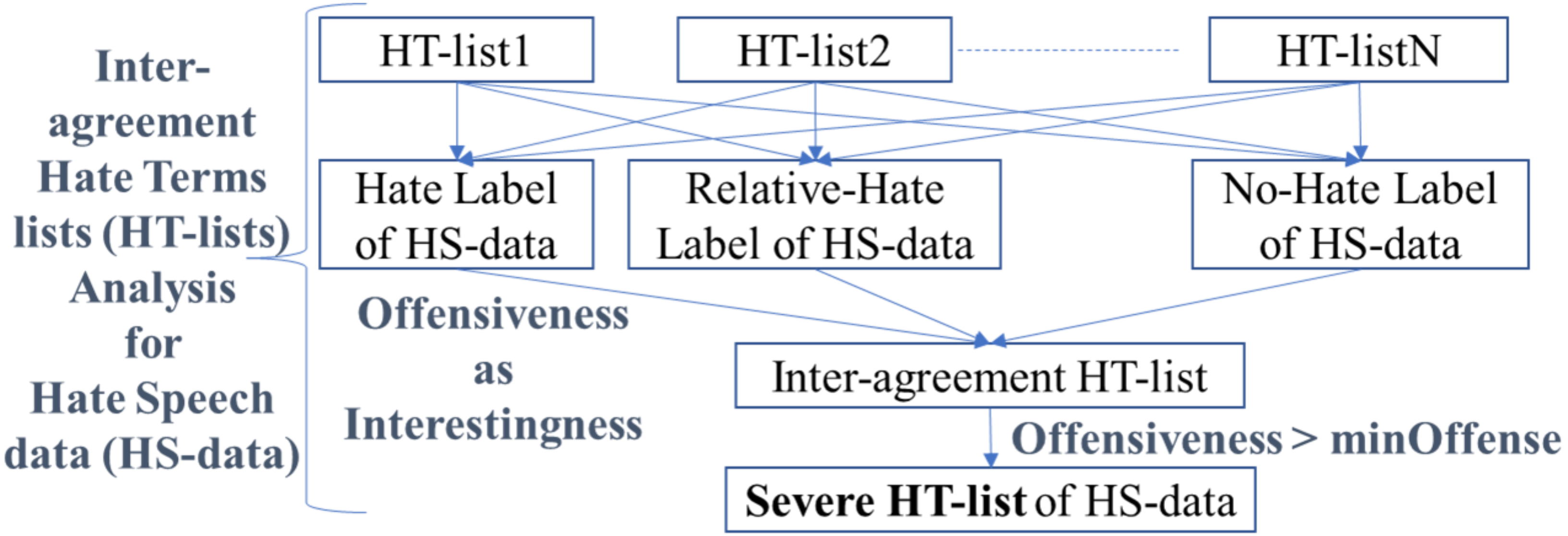
- Percentage contribution (i.e., overall input towards cost) of a hate term occurrences to a hate class.
- Varying value of the Offensiveness of a HT for a given HS-data divided into classes (Hate, Relative-hate, and No-hate).
- HT will be most Hateful when its Offensiveness equals to 1
- HT will be least hateful when its Offensiveness value is equal to 0

$$\text{Offensiveness} = \frac{2 \times \text{Hatefulness} \times \text{Relativeness}}{(\text{Hatefulness} + \text{Relativeness})}$$



Severe Hate Terms-list (Answers to RQ2)

- Generate the Severe HTs-list from the Inter-Agreement HTs-list having HTs with ***Offensiveness* metric** values greater than a user-defined interestingness threshold ***minimum Offense*** (**minOffense**).
- Offensiveness provides help to separate out the highly severe HTs and the less severe HTs.
- High values of Offensiveness generate the Severe HTs-lists.
- Severe HTs-list helps in better hate speech classification as compared to the given set of HTs-lists.



Inter-agreement Confusion-matrix (7th Artifact)

- Information about confusion-matrix with
  - True Positive (TP),
  - True Negative (TN),
  - False Positive (FP), and
  - False Negative (FN)
- For the calculation of accuracy, precision, recall, and f-measure of HS classification.
- To avoiding **imbalance**: Percentage of HS-lines in a class to evaluate metrics

CaseStudy class	**TP** = percentage of HS-lines	**TN** = percentage of HS-lines	**FP** = percentage of HS-lines	**FN** = percentage of HS-lines
Hate	with HTs occurring in Hate class	without HTs occurring in NonOffensive class	with HTs in NonOffensive class	without HTs in Hate class
Relative-hate	with HTs occurring in Offensive class	without HTs occurring in NonOffensive class	with HTs in NonOffensive class	without HTs in Offensive class
Hate + Relative-hate	with HTs occurring in Hate+Offensive class	without HTs occurring in NonOffensive class	with HTs in NonOffensive class	without HTs in Hate+Offensive class

Summary_N(HateTerms) (8th Artifact)

- This provides information of percent HS-lines with N HTs in a HS-data class e.g., x\% have 1 HT, y\% have 2 HTs, z\% have 3 HTs and so on.

Discussed with experiments

Rare instances of the co-occurring HTs

- Imbalance occurrences of hate speech as compared to normal speech leads to rare instances of HTs and HS-lines in a HS-data.
- Identify and list those rare HTs by identifying rare concepts and their effect on the classes.
- It is interesting to analyse those groups of rare HTs (as hate concepts) and their effect on the classes.

Qualitative analysis:
Stable Hate Rules, Concepts, Transitivities, and Lattices

1. Quantitative Analysis

- To create an *Inter-agreement HTs-list*, which explains the contribution of an individual hate term toward hate speech.
- To produce a Severe Hate Terms list (*Severe HTs-list*)

2. Qualitatively Analysis

- ***Stable Hate Rule* (*SHR*) mining detects ordered frequently co-occurring HTs with *minimum Stability* (*minStab*).This form *Stable Hate Rules* and *Concepts*.**
- **These rules and concepts are used to visualise the graphs of *Transitivities* and *Lattices* formed by HTs.**

Interestingness thresholds

- It uses multiple thresholds to retrieve interesting and significant rules
- It separates interesting rules from the less or non interesting rules
- **A** and **B** together (where **A → B**) can have three interestingness thresholds:

  1) *minimum Support* (***minSup***) is a threshold for minimum number of occurrences of HTs **A** and **B** occurring together,

  2) *minimum Confidence* (***minConf***) is a threshold for minimum number of occurrences of **A ∪ B** divided by number of occurrences of HT **A** i.e., $\mathbf{N(A \cup B) \div N(A)}$.

  3) *minimum Stability* (***minStab***) [5][7] is a threshold for minimum number of states in which rule exceeds minSup & minConf

[5] A. Chaturvedi, A. Tiwari, and N. Spyratos. "minStab: Stable Network Evolution Rule Mining for System Changeability Analysis." *IEEE Trans. on Emerging Topics in Computational Intelligence* (2019).
[7] A. Chaturvedi, A. Tiwari, and N. Spyratos. "System Network Analytics: Evolution and Stable Rules of a State Series." *IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2022.

Hate Speech Rule Mining Example

- To discover co-occurrences of desired terms
  - consider only HTs and contextual terms in a hate speech
- Suppose 'Anglo' is a contextual term, which is a white English speaking person.
- Suppose there are 19 tweets (each as a hate speech) with `sp*c', which is an ethnic slur for people from Spanish-speaking.
- Out of them 3 tweets are as follows
  - **Tweet 1:** "Black cops k*ll white citizens. sp*c cops k*ll Anglo citizens. Z*geuner cops r*pists."
  - **Tweet 2:** "No half-breed sp*c Anglo, k*lled so."
  - **Tweet 3:** "A*glo-S*xn Protestant, alive US. None, foreign f*lth."

Hate Speech Rule Mining Example

- The FreqHTs denotes the frequency of a Hate Term (HT) (means number of occurrences of individual HT) in a hate speech.
- The FreqHT of 'Anglo' and 'sp*c' are as follows: N(Anglo) = 3 and N(sp*c) = 18.
- The FreqCoHTs denote the frequency of co-occurring HTs in a hate speech.
- The FreqCoHTs for (Anglo and sp*c) are as follows: N(Anglo, sp*c) = 2; N(Anglo as antecedent) = 1; and N(sp*c as antecedent) = 15.

[Anglo] $\rightarrow$ [sp*c] #SUP:2 #CONF: 0.66 means
N(Anglo $\cup$ sp*c) / N(Anglo) = 2/3
[sp*c] $\rightarrow$ [Anglo] #SUP: 2 #CONF: 0.11 means
N(Anglo $\cup$ sp*c) / N(sp*c) = 2/18

Treat as unordered database result in the following unordered hate rules

Ordered sequence database result in the following ordered hate rule

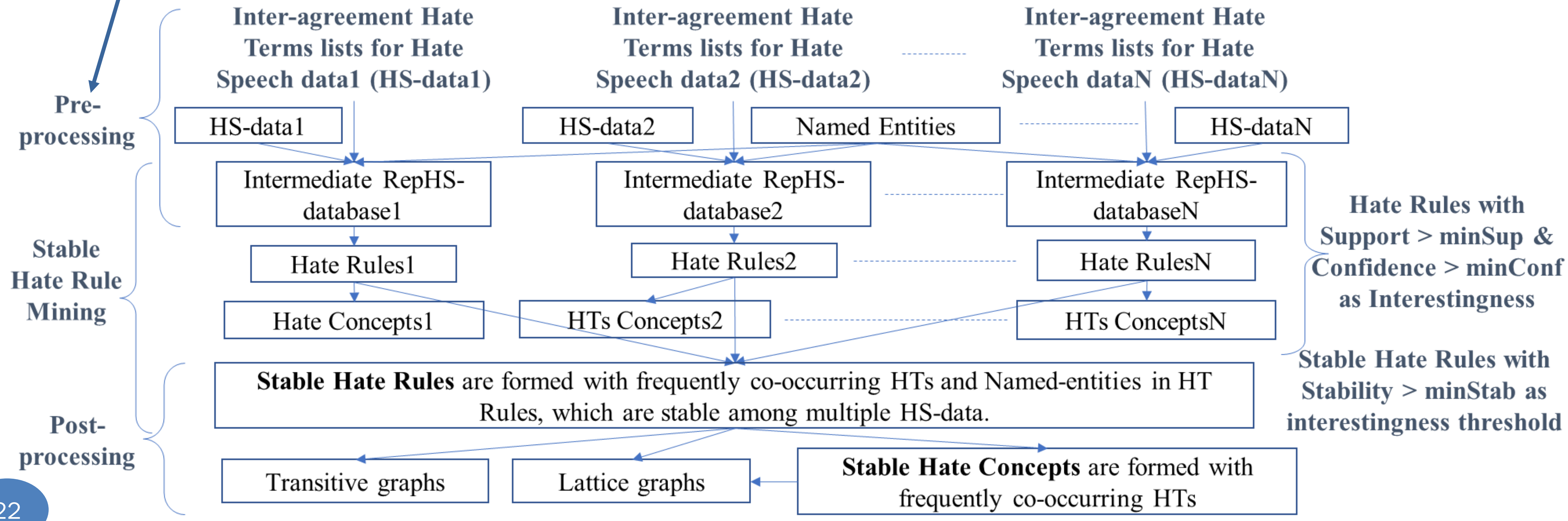
[sp*c] $\rightarrow$ [Anglo] #SUP: 2 #CONF: 0.13 means
N(Anglo $\cup$ sp*c) / N(sp*c as antecedent) = 2/15

Stable Hate Rule (SHR)

- The **stability** is the number of HS-data in which a *hate rule* occurs with sufficient minSup and minConf.
- Hate rule occurring more than a **minStab** number are said to be *Stable Hate Rule*.
- SHR mining is performed over multiple Hate Speech data (HS-data) with only hate terms and Named-entities.
- This generated Stable Hate Rules (SHRs), which can be read as "if someone uses a HT 'A', then most probably the person may also use HT 'B' with a given probability".
- The SHRs could be like [A] → [B], where the [A] is antecedent and the [B] is its consequent.

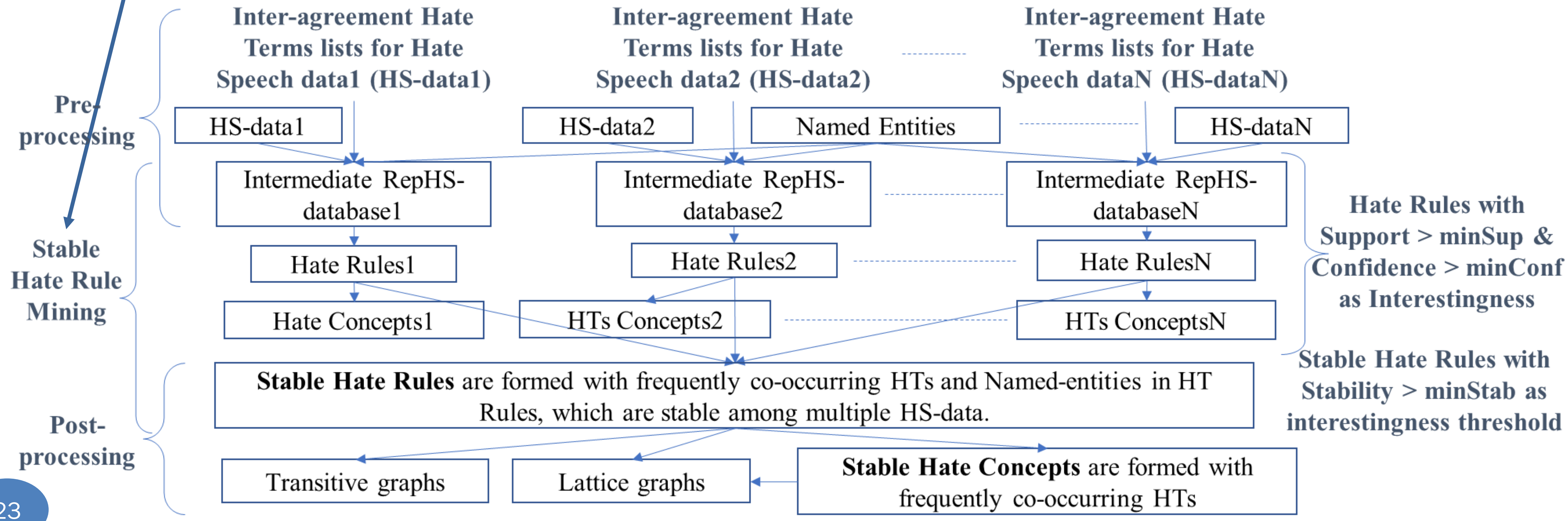
Stable Hate Rule (SHR) (Answers to RQ4a)

- **Pre-processing:**
  - Inter-agreement HTs-list is used to make an Representational Hate Speech Database (RepHS-database)



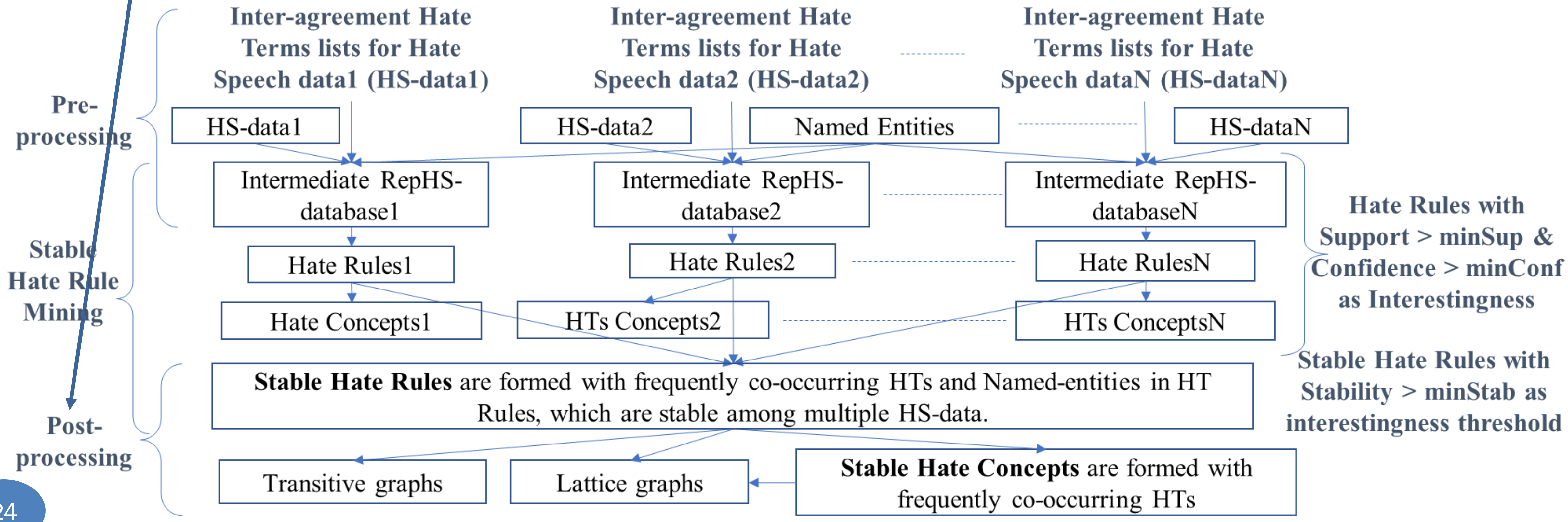
Stable Hate Rule (SHR) (Answers to RQ4a)

- **Stable Hate Rule (SHR) mining**
  - SHR mining over the database to discover co-occurring HTs.
  - This helps to discover and analyse the co-occurring concepts of HTs.



Stable Hate Rule (SHR) (Answers to RQ4a)

- **Post-processing** visualization as Transitive graph and Lattice graph
  - Both visualizes hate rules with similar Hate-Terms by forming graphs



Dataset

Hate Speech data (HS-data)

Hate Terms-lists (HTs-lists)

Hate Speech data (HS-data)

Three hate speech datasets and six hate terms lists.

a) Davidson et al. [8] (Twitter tweets)

b) de Gibert et al. [10] (White Supremacy forum)

c) Gao et al. [11] (Fox-news-comments)

Hate Speech data	**Classes**	
	Used in HS-data	***Used in our work***
Davidson et al. [8]	Hate	Hate
	Offensive	Relative-Hate
	Non-Offensive	No-Hate
de Gibert et al. [10]	Hate	Hate
	Relational Hate	Relative-Hate
	No-Hate	No-Hate
Gao et al. [11]	Hate	Hate
	–	Relative-Hate
	No-Hate	No-Hate

[8] T. Davidson, et al. "Automated hate speech detection and the problem of offensive language." Int. AAAI Conf. on Web and Social Media. Vol. 11. No. 1. 2017.
[10] O. de Gibert, et al. "Hate speech dataset from a white supremacy forum." arXiv preprint arXiv:1809.04444 (2018).
[11] L. Gao, and R. Huang. "Detecting online hate speech using context aware models." arXiv preprint arXiv:1710.07395 (2017).

Hate Terms-lists (HTs-lists)

a) Chandrasekharan et al. [12] contains Reddit hate lexicon[1]
b) Gorrell et al. [13] contains abuse lexicon in tweets related to UK politicians[2]
c) Hatebase[3] contains a various kinds of hate vocabulary from many countries

[12] E. Chandrasekharan, et al. "You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech." *ACM on Human-Computer Interaction* 1.CSCW (2017): 1-22.
[13] G. Gorrell, et al. "Twits, twats and twaddle: trends in online abuse towards UK politicians." *Int. AAAI Conf. on Web and Social Media*. Vol. 12. No. 1. 2018.
1 https://www.dropbox.com/sh/5ud4fwxvb6q7k20/AAAH SN8i5cfmJRKJteEW2b2a
2 https://cloud.gate.ac.uk/shopfront/displayItem/gate-hate
3 https://hatebase.org/academia

Hate Terms-lists (HTs-lists)

d) Bassignana et al. [14] list named Hurtlex[4] contains lexicons of hate terms for 50 languages, which are divided into 17 categories.

e) Wiegand et al. [15] filtered abusive words from negative polar expressions[5].

f) Union: We made a union list from all the distinct HTs

[14] E. Bassignana, V. Basile, and V. Patti. ``Hurtlex: A multilingual lexicon of words to hurt." *5th Italian Conf. on Computational Linguistics, CLiC-it* 2018. Vol. 2253. CEUR-WS, 2018.

[15] M. Wiegand, et al. "Inducing a lexicon of abusive words–a feature-based approach." *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long Papers). 2018.

4 https://github.com/valeriobasile/hurtlex

5 https://github.com/uds-lsv/lexicon-of-abusive-words

Hate Speech Analytics and Experiments

A. Generation of Severe Hate Terms List

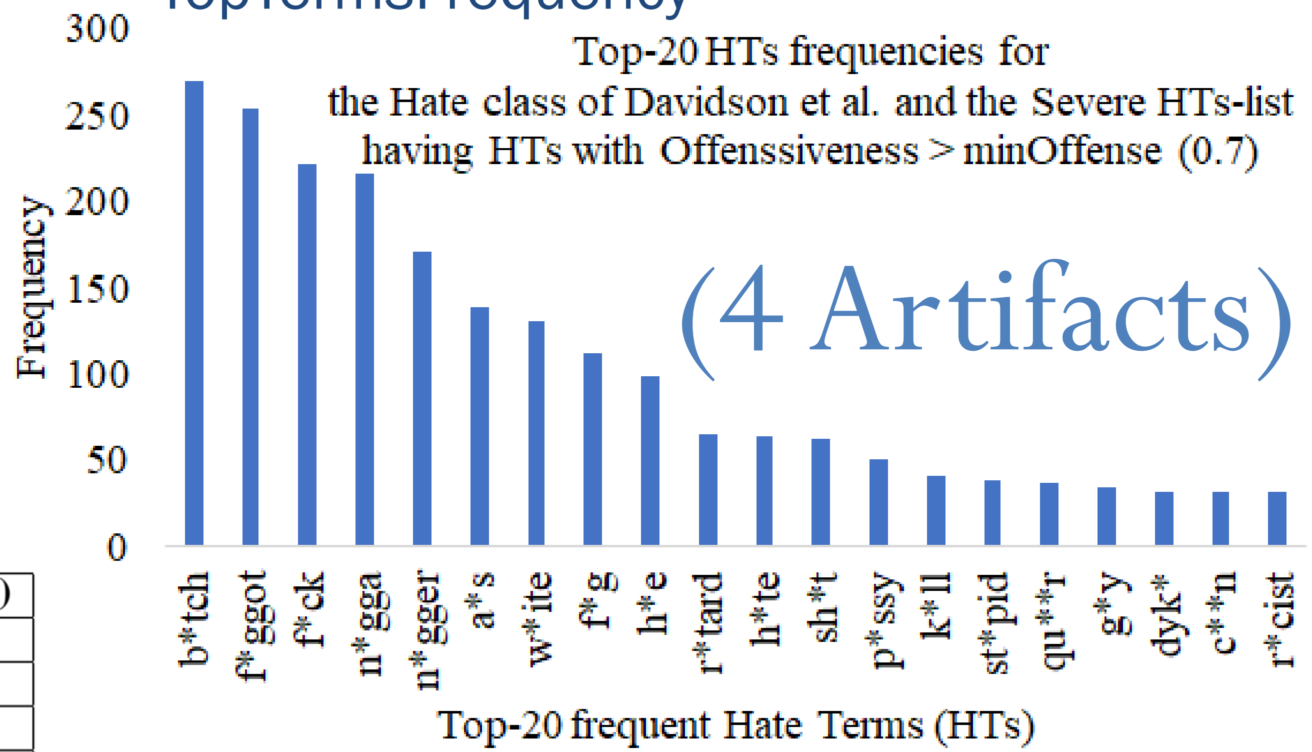
B. Stable Hate Rules, Concepts, Transitivities, and Lattices

1. Creation of hate terms frequencies

N(0), N(1), N(2) ... N(X) Terms Example.

Filename	Hate Term	Tweets
N(0)_HTs	–	#[IDENTITY] can get a job at the [IDENTITY]. Or as The [IDENTITY]. I hear they like diversity and tolerance. As long as you ain't a cracker #[TAG]
N(1)_HTs	f*ggot	@[IDENTITY] answer my [IDENTITY] f*ggot #[TAG]
N(2)_HTs	f*ggot; f*ck	@[IDENTITY] f*ck those f*ggots
so on ...	...	...

2. AllHateTermsFrequencies and TopTermsFrequency



3. AllHTsPercentLine

Hate Term	N(HateTermInLines)	N(Lines)	%(HateTermLines)
f*ggot	249	1430	17.413
b*tch	240	1430	16.783
f*ck	199	1430	13.916
so on	...	...	...

OuterJoinHTsFrequencies Example.

Hate Term	Davidson et al. 0Hate	Davidson et al 1Offensive	Davidson et al. 2NonOffensive
f*ggot	253	291	1
b*tch	269	11192	11
f*ck	221	2039	–
so on	...	...	...

OuterJoinHTsPercentLines Example.

Hate Terms	Davidson et al. 0Hate	Davidson et al 1Offensive	Davidson et al. 2NonOffensive
f*ggot	17.413	1.501	0.024
b*tch	16.783	54.627	0.264
f*ck	13.916	9.734	–
so on	...	...	...

4. OuterJoinHTsFrequencies and OuterJoinHTsPercentLines

Intra-Agreement-HTs for each HTs-list (5th Artifact)

INTRA-AGREEMENT HTS EXAMPLE FOR HS-DATA (DAVIDSON ET AL.) AND HTS-LIST (UNION).

Hate Terms (HTs)	**Hate Class HS-lines**	**#Offensive + Non-Offensive HS-lines**	**#Hate Class HS-lines**	**Hatefulness (Hate Class)**	**Relativeness (Hate Class)**	**#Hate + Offensive HS-lines**	**Non-Offensive HS-lines**	**#Hate + Offensive HS-lines**	**Hatefulness (Hate + Offensive)**	**Relativeness (Hate + Offensive)**
f*ggot	249	1	1431	1	0.996	537	1	20622	1	0.998
b*tch	240	11	1431	1	0.956	10723	11	20622	1	0.999
f*ck	199	0	1431	1	1	2067	0	20622	1	1
tr*sh	106	680	1431	1	0.135	442	680	20622	1	0.394
eurotr*sh	0	1	1431	0	0	1	1	20622	1	0.5
tr**ler park tr*sh	2	1	1431	1	0.667	2	1	20622	1	0.667
tr**ler tr*sh	3	2	1431	1	0.6	6	2	20622	1	0.75
white tr*sh	56	3	1431	1	0.949	91	3	20622	1	0.968
so on ...	...	...	...	...	...	...	...	...	...	...

Inter-Agreement-HTs for multiple HTs-lists (6th Artifact)

INTER-AGREEMENT HTS BETWEEN THE DAVIDSON ET AL. AND THE SIX GIVEN HTS-LISTS.

HTs	Hatefulness (Hate)	Relativeness (Hate)	Offensiveness (Hate)	Hatefulness (Hate+Offensive)	Relativeness (Hate+Offensive)	Offensiveness (Hate+Offensive)	HateListNames
f*ggot	1	0.996	0.998	1	0.998	0.999	Chandrasekharan et al Reddit hate lexicon; Gorrell et al abuse-terms; HateBaseList; hurtlex_EN; Union; Wiegand et al
b*tch	1	0.956	0.978	1	0.999	0.999	Gorrell et al abuse-terms; HateBaseList; hurtlex_EN; Union;
f*ck	1	1	1	1	1	1	hurtlex_EN; Union; Wiegand et al
tr*sh	1	0.135	0.238	1	0.394	0.565	HateBaseList; hurtlex_EN; Union
eurotr*sh	0	0	NaN	1	0.5	0.667	HateBaseList; Union
tr**ler park tr*sh	1	0.667	0.8	1	0.667	0.8	HateBaseList; Union
tr**ler tr*sh	1	0.6	0.75	1	0.75	0.857	HateBaseList; Union
white tr*sh	1	0.949	0.974	1	0.968	0.984	HateBaseList; Union
so on ...	...	...	...	...	...	...	...

Answer to RQ2:

Answers to RQ3:

Inter-agreement Confusion-matrix (7th Artifact)

FOR THE THREE HS-DATA, THE TABLE PROVIDES A COMPARISON OF THE SEVERE HTS-LIST WITH THE GIVEN HTS-LISTS.

HTs-list Name (minOffense, number of HTs)	**HS-data Name and Class**	**Accuracy**	**Recall**	**Precision**	**F-Measure**	**Compute Time**
Gorrell et al abuse-terms (-, 403)	Davidson_et_al_ 0Hate Vs. No-Hate	0.857	0.784	0.917	0.845	12 sec
	Davidson_et_al_ 0Hate+1Offensive Vs. No-Hate	0.845	0.761	0.915	0.831	
	Davidson_et_al_1 Offensive Vs. No-Hate	0.844	0.759	0.915	0.83	
Offensiveness(Hate) (0.7, 298)	Davidson_et_al_ 0Hate Vs. No-Hate	**0.921**	**0.946**	0.901	**0.923**	17 sec
	Davidson_et_al_ 0Hate+ 1Offensive Vs. No-Hate	**0.929**	**0.962**	0.903	**0.931**	
	Davidson_et_al_1 Offensive Vs. No-Hate	**0.93**	**0.963**	**0.903**	**0.932**	
Union (-, 13538)	de_Gibert_et_al_ 0Hate Vs. No-Hate	0.633	0.959	0.58	0.723	1 min 31 sec
	de_Gibert_et_al_0Hate +1RelationalHate Vs. No-Hate	0.629	0.951	0.578	0.719	
	de_Gibert_et_al_ 1RelationalHate Vs. No-Hate	0.6	0.893	0.563	0.69	
Offensiveness(Hate) (0.46, 578)	de_Gibert_et_al_ 0Hate Vs. No-Hate	**0.821**	0.832	**0.814**	**0.823**	**14 sec**
	de_Gibert_et_al_ 0Hate+ 1RelationalHate Vs. No-Hate	**0.8**	0.789	**0.806**	**0.797**	
	de_Gibert_et_al_ 1RelationalHate Vs. No-Hate	**0.646**	0.482	**0.718**	0.577	
Union (-, 13538)	Gao_et_al_ 0Hate Vs. No-Hate	0.46	0.772	0.475	0.588	15 sec
Offensiveness(Hate) (0.75, 622)	Gao_et_al_ 0Hate Vs. No-Hate	**0.541**	0.718	**0.53**	**0.61**	**5 sec**

Our approach shown an improvement from 0.845 to 0.923 (best) as compared to the baseline.

Severe HTs-list provides better results for confusion-matrix (precision, recall, f-measure, and accuracy).

Answers to RQ3:

- Two facts for a HS-data.
  - **Fact 1:** for best recall, the FN should be zero.

    This happens when all HTs (in a HTs-list) are found in the Hate class of HS-data.
    - Example, a large HTs-list tends to a low FN.
  - **Fact 2:** for best precision, the FP is zero.

    This happens when no HTs (in a HTs-list) are found in the No-Hate class of HS-data.
    - Example, a small HTs-list tends to a low FP.
- The best conditions to select HTs leads to best precision and recall,

thus we can generate a Severe HTs-list.

RANKING OF HTS-LIST NAME IN DECREASING ORDER OF INTER-AGREEMENT WITH THE HS-DATA.

HS-data Name	HTs-lists Names (number of HTs)
Davidson et al	**Offensiveness(Hate) (0.7, 298)**, Gorrell et al abuse-terms (403), HateBaseList (1015), Wiegand et al lexicon-of-abusive-words (7156), Hurtlex EN (5925), Union (13538), and Chandrasekharan et al Reddit hate lexicon (199).
de Gibert et al	**Offensiveness(Hate)(0.46, 578)**, Union (13538), Hurtlex EN (5925), Wiegand et al lexicon-of-abusive-words (7156), Chandrasekharan et al Reddit hate lexicon (199), HateBaseList (1015), Gorrell et al abuse-terms (403).
Gao et al	**Offensiveness(Hate)(0.75, 622)**, Union(13538), Wiegand et al lexicon-of-abusive-words (7156), Hurtlex EN (5925), Chandrasekharan et al Reddit hate lexicon (199), Gorrell et al abuse-terms (403), HateBaseList(1015).

Summary N(HateTerms) (8^{th} Artifact)

For the three HS-data and six HTs-list, the table provide summarised overview.

Dataset Name and Class	HateList Name	HateTerms(N)	N(Entries)	TotalLines	%(Entries)
Davidson et al 0Hate	Chandrasekharan et al Reddit hate lexicon	0	581	1430	40.629
Davidson et al 0Hate	Chandrasekharan et al Reddit hate lexicon	1	671	1430	46.923
so on ...	...	...	...	...	...
Davidson et al 1Offensive	Chandrasekharan et al Reddit hate lexicon	0	16101	19190	83.903
Davidson et al 1Offensive	Chandrasekharan et al Reddit hate lexicon	1	2654	19190	13.83
so on ...	...	...	...	...	...

Hate Speech Analytics and Experiments

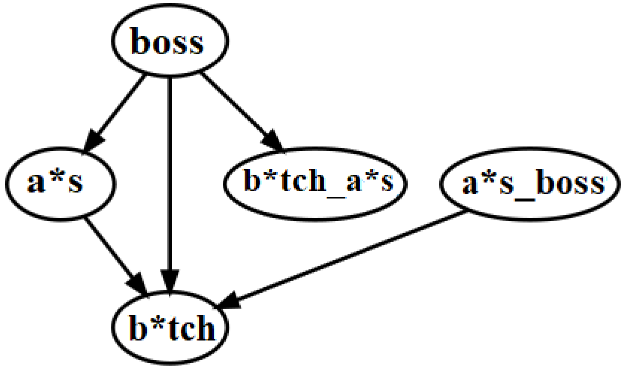
A. Generation of Severe Hate Terms List

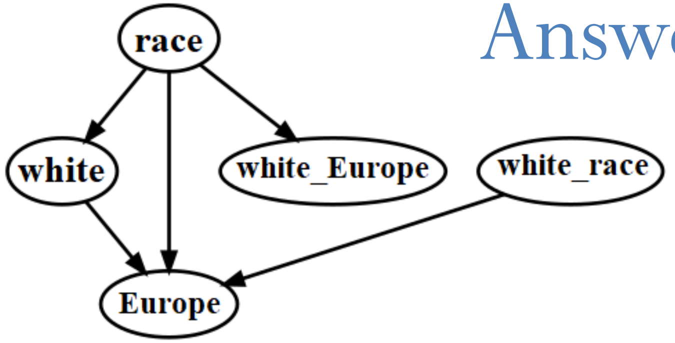
B. Stable Hate Rules, Concepts, Transitivities, and Lattices

SHR mining to generate: Stable Hate Rules, Concepts, Transitivities, and Lattices

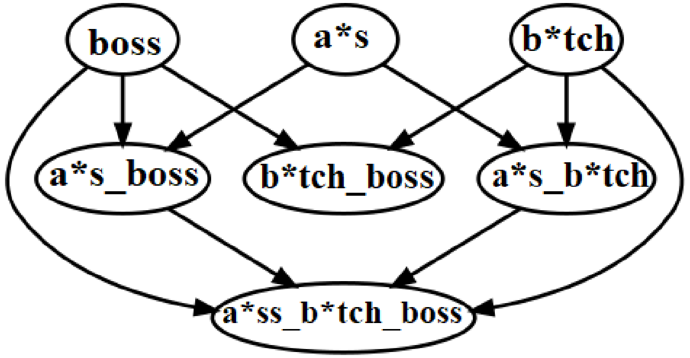
TWO HATE CONCEPTS (FIRST ROW) AND THEIR SHRS WITH SIMILAR HTS.

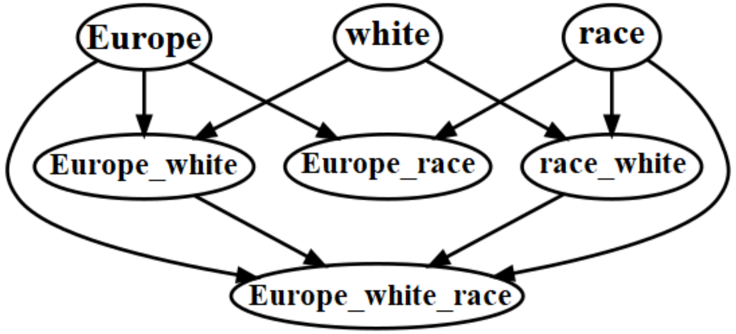
a*s b*tch boss 5	**Europe race white 5**
a*s $\rightarrow$ b*tch	white $\rightarrow$ Europe
boss $\rightarrow$ b*tch a*s	race $\rightarrow$ white Europe
a*s boss $\rightarrow$ b*tch	white race $\rightarrow$ Europe
boss $\rightarrow$ b*tch	race $\rightarrow$ white
boss $\rightarrow$ a*s	race $\rightarrow$ Europe





Answers to RQ4b





Conclusions

Conclusions

- To collect Inter-agreement information about the HTs-list (Hate Terms list) and the HS-data (Hate Speech data),
  - answered the four research questions.
- Generated reports that include: top frequent HTs, intra/inter-agreement of HTs in HTs-list with the HS-data, summarized hate-term occurrences, and Offensiveness of HTs.
- For quantitative analysis,
  - proposed threshold minOffense for HTs,
  - our Severe HTs-list has out-performed all the given HTs-lists.
- For qualitative analysis,
  - our SHRs provided visual analytic as Transitive and Lattice graphs of the HTs co-occurring in HS-data for context of Women and Regions.

Acknowledgment

Acknowledgment

- Thanks to
  - *Prof. Nishanth Sastry* (University of Surrey)
  - *Dr. Bertie Vidgen* (The Alan Turing Institute)
  - *Dr. Jatinder Singh* (Cambridge University)
- Also thanks for fellowship to Dr. Animesh Chaturvedi as Post Doctoral
  - *The Alan Turing Institute* (U.K.)
  - *King's College London* (U.K.)

Related Publications

Citation:

Animesh Chaturvedi, and Rajesh Sharma
"minOffense: Inter-Agreement Hate Terms for Stable Rules, Concepts, Transitivities, and Lattices"
IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2022.

Stable Rule Mining

- A. Chaturvedi and A. Tiwari. “System Evolution Analytics: Evolution and Change Pattern Mining of Inter-Connected Entities”. *IEEE International Conference on Systems, Man, and Cybernetics (SMC)* 2018.
- A. Chaturvedi, A. Tiwari, and N. Spyratos “minStab: Stable Network Evolution Rule Mining for System Changeability Analysis”. *IEEE Trans. on Emerging Topics in Computational Intelligence*, 2019.
- A. Chaturvedi, A. Tiwari, and N. Spyratos. “System Network Analytics: Evolution and Stable Rules of a State Series.” *IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2022.

ขอบคุณ
Thai

Grazie
Italian

תודה רבה
Hebrew

धन्यवादः
Sanskrit

ಧನ್ಯವಾದಗಳು
Kannada

Ευχαριστώ
Greek

Thank You
English

Gracias
Spanish

Спасибо
Russian

Obrigado
Portuguese

شكراً
Arabic

https://sites.google.com/site/animeshchaturvedi07

Merci
French

多謝
Traditional Chinese

धन्यवाद
Hindi

Danke
German

多谢
Simplified Chinese

நன்றி
Tamil
Tamil

ありがとうございました
Japanese

감사합니다
Korean