Assignment 2 : Annotation
Team - GR-9-AI
Tanya Kaintura, Payanshi Jain, Rishikesh Ramachandran, Antonios Georgakopoulos

---

Subtask 2:

Inter-annotator agreement calculation:

- Cohen's Kappa:
  It is a measurement used to assess the level of agreement between two or more annotators when categorising or labelling items. It is given by:

$$k = \frac{p_0 - p_e}{1 - P_e}$$

  Where, p_0 is the observed proportion of agreement between annotators and p_e is the
  Expected proportion of agreement by chance.

- **Percentage Agreement:**
  Percentage agreement is a simple measure that calculates the agreement between annotators as a percentage of the judgments they both agree on, out of the total judgments made. It is given by,

$$Percentage\ Agreement\ = \frac{Number\ of\ Agreements}{Total\ Number\ of\ Judgments} * 100$$

Inter-annotator agreement calculations for l_Gibert(HATE/NOHATE):

Contingency table:

	HATE	**NOHATE**	**ALL**
HATE	11	2	13
NOHATE	4	27	31
ALL	15	29	44

Cohen's Kappa value: 0.686
Percentage Agreement value: 86.36%

Inter-annotator agreement for l_Kumar(CAG/NAG/OAG):

Contingency table:

	CAG	**NAG**	**OAG**	**ALL**
CAG	4	4	11	19
NAG	2	10	0	12
OAG	3	1	9	13
ALL	9	15	20	44

Cohen's Kappa value: 0.302
Percentage Agreement value: 52.27%

Inter-annotator agreement calculations for l_Zamp(OFF/NON):

Contingency table:

	NON	OFF	ALL
NON	19	2	21
OFF	2	21	23
ALL	21	23	44

Cohen's Kappa value: 0.818
Percentage Agreement value: 90.91%

Inter-annotator agreement calculations for l_Zamp(TARG/NOTARG):
Contingency table:

	-	NOTARG	TARG	ALL
-	19	1	1	21
NOTARG	0	1	2	3
TARG	2	2	16	20
ALL	21	4	19	44

Cohen's Kappa value: 0.681
Percentage Agreement value: 81.82%

Inter-annotator agreement calculations for l_Zamp(G/I/O):
Contingency table:

	-	G	I	O	ALL
-	21	2	1	0	24
G	0	10	0	0	10
I	2	2	4	0	8
O	2	0	0	0	2
ALL	25	14	5	0	44

Cohen's Kappa value: 0.657
Percentage Agreement value: 79.54%

CONFUSION MATRIX:

		Predicted	
		HATE	NOHATE
Actual	**HATE**	11	2
	NOHATE	4	27

HATE vs. NOHATE: Both annotators largely agree when categorizing 'NOHATE' instances (24 times), but there are differences in identifying 'HATE' (10 agreed instances, 3 disagreements). The 7 instances where one annotator identified 'HATE' and the other 'NOHATE' suggest that the criteria for identifying hate speech may not be fully consistent between the two.

		Predicted		
		CAG	**OAG**	**NAG**
Actual	**CAG**	4	11	4
	OAG	3	9	1
	NAG	2	0	10

CAG, OAG, NAG: There's considerable disagreement here, especially when classifying 'CAG' and 'OAG'. This could indicate that the distinction between these categories is not very clear to the annotators. The agreement is highest in the 'NAG' category with 9 instances, indicating better clarity for this category.

		Predicted	
		OFF	NON
Actual	**OFF**	21	2
	NON	2	14

OFF vs. NON: This matrix shows strong agreement between the annotators, especially in 'NON' (18 times). The 'OFF' category has 21 agreements and just 2 disagreements, suggesting a high level of consensus on what constitutes offensive content.

		Predicted		
		TARG	**NOTARG**	**-**
Actual	**TARG**	16	2	2
	NOTARG	2	1	0
	-	1	1	19

TARG vs. NOTARG: There is a fair amount of agreement here, especially in the 'TARG' category (15 times). The '-' category is for where Annotator 1 has annotated the sentence as NON in LevelA, these cases has also been captured here in the table, has the highest agreement with 18 instances.

		Predicted			
		-	**G**	**I**	**O**
Actual	**-**	21	2	1	0
	G	0	10	0	0
	I	2	2	4	0
	O	2	0	0	0

G, I, O: The annotators largely agree on 'G' and '-', but there is some disagreement in the 'I' category. This suggests that there might be some ambiguity or subjective interpretation in what falls under 'I'.

Overall, the matrices suggest a moderate to high level of agreement between the two annotators, depending on the category. However, there are still several instances of disagreement that might require further discussion or clarification of the annotation guidelines.

Error Analysis

1. Given the confusion matrix above for HATE/NOHATE, majority of the time Annotator 1 has annotated as NOHATE and annotator 2 has annotated HATE for that sentence. This can be a systematic error based on the understanding of the guidelines or different understanding of the context as there is a clear pattern of error in this case.

Example cases -

	l_Gibert(HATE/NOHATE)	
text	**Annotator 1**	**Annotator 2**
False news	NOHATE	HATE
He is coward	NOHATE	HATE
In JNU students is learning from @ArvindKejriwal how to speak from a negative.. Shame on you kejri sir no reaction on it #ShutDownJNU	NOHATE	HATE
@USER @USER @USER Bullshit! Regulations are good because it keeps conservatives from striping our land into the Abyss	NOHATE	HATE
Confirm the MAN! THE LEFT WING ACTIVIST IS A LIAR WHO'S STORY CHANGES! NO HARD EVIDENCE! #ChristineBlaseyFord Innocent until proven guilty! #ConfirmKavanaugh #MAGA @USER #WeakRinos #Republicans	NOHATE	HATE
I'm Assuming I'm Not Going to Get an Answer': Brainless #Liberal Stooge Ocasio-Cortez .@USER on $40T Policy Plan URL #WakeUpAmerica THIS is NOT America's future VOTE Republican or surrender America to these #DEM morons @USER #MAGA	NOHATE	HATE

2. Most disagreements are lying in annotations for l_Kumar(CAG/NAG/OAG) . Even based on the contingency table the percentage agreement is 52.27%. As shown in the confusion table matrix the majority confusion lies in the annotating CAG and OAG. As there is no clear pattern this could be a misinterpretation of guidelines, possibly due to differences in interpreting what constitutes "Covert Aggression" (CAG) or "Non-Aggressive" (NAG). it can also be that annotators have different threshold for identifying covert aggression.

Example cases -

	l_Zamp(OFF/NON)	
text	**Annotator 1**	**Annotator 2**
4 people died and Indian express finds humour in it..rip journalism..	CAG	NAG
False news	NAG	OAG
He is coward	CAG	OAG
In JNU students is learning from @ArvindKejriwal how to speak from a negative.. Shame on you kejri sir no reaction on it #ShutDownJNU	CAG	OAG
our President has no history like modi who kills thousand Muslims in gujrat	OAG	CAG
People of Kerala will give befitting reply to these BJP idiots. They Are wise than other northern States	CAG	OAG
Same Pakistani mentality	OAG	CAG
This is disgusting	CAG	NAG

3. Looking at the confusion matrix for l_Zamp(OFF/NON) there are only 5 disagreements and 90% percentage agreement based on contingency tables. In these instances, both annotators were in concordance, classifying the texts consistently either as offensive or non-offensive as also seen in examples given below. This strong agreement suggests that the annotators found the presence or absence of offensive language in these texts to be self-evident, leaving little room for interpretation or debate. It underscores the clarity in identifying offensive content when it is unequivocally present, highlighting the efficacy of the annotation guidelines and the shared understanding between annotators in discerning the offensive nature of language in these particular instances. This level of agreement is pivotal for establishing a reliable and consistent annotation process in tasks related to offensive language detection.

Example cases

	l_Zamp(OFF/NON)	
text	**Annotator 1**	**Annotator 2**
#Dutch people who live outside of #NewYorkCity are all white trash.	OFF	OFF

4 people died and Indian express finds humour in it..rip journalism..	NON	NON
He is coward	OFF	OFF
Its not good signs....Webcast crashed...CNBC tvs 18 webcast playing like cat and mouse...awefull songs	OFF	OFF

To summarize, the errors in the "HATE/NOHATE" category appear systematic, with Annotator 1 predominantly annotating as "NOHATE" and Annotator 2 as "HATE." This could be due to differences in interpreting hate speech or contextual nuances. In contrast, the "l_Kumar(CAG/NAG/OAG)" category exhibits more diverse errors, with no discernible pattern. The "l_Zamp(OFF/NON)" category demonstrates a high level of agreement, indicating a clear distinction between offensive and non-offensive content. These findings emphasize the importance of well-defined annotation guidelines and consistency in interpretation when conducting tasks related to offensive language detection.

The Differences between the sets of guidelines:

Do they have similar definitions of hate speech? Do they address the same phenomenon?

All three guidelines share a similar definition of hate speech: any form of communication that denigrates a specific group of people based on characteristics such as race, colour, ethnicity, sexual orientation, nationality, religion, or other attributes. They each either fully or partially address the same phenomenon.

For instance, in the paper titled "Hate Speech Dataset from a White Supremacy Forum" by O. de Gibert et al. (2018), the guidelines in the paper aim to address the problem of online hate speech and the growing need for automated detection of such harmful content. Similarly, the paper titled "Predicting the Type of Offensive Posts in Social Media" by Marcos Zampieri et al. (2019) seeks to address the prevalence of abusive language on social media platforms, including hate speech, cyberbullying, and cyber-aggression.

Can they be reliably annotated?

Based on the results of the Cohen's Kappa test and Percentage Agreement test, it has been determined that the three-level hierarchical annotation schema presented in the paper "Predicting the Type of Offensive Posts in Social Media" by Marcos Zampieri et al. (2019) and the annotation schema outlined in the paper "Hate Speech Dataset from a White Supremacy Forum" by O. de Gibert et al. (2018) exhibit a high level of reliability. This conclusion is drawn from the significantly high values obtained in both of these evaluation tests.

Conversely, the annotation schema provided in the paper "Aggression-annotated Corpus of Hindi-English Code-mixed Data" by Ritesh Kumar et al. (2018) is considered less reliable due to the lower values observed in both the Cohen's Kappa test and Percentage Agreement test. According to the authors of this paper, these lower values may be due to the broader scope of interpretation allowed for annotators.

Are the guidelines clear and do they cover all cases?

The papers "Hate Speech Dataset from a White Supremacy Forum" by O. de Gibert et al. (2018) and "Predicting the Type of Offensive Posts in Social Media" by Marcos Zampieri et al. (2019) offer clear guidelines and comprehensively address various scenarios. However, the paper "Aggression-annotated Corpus of Hindi-English Code-mixed Data" by Ritesh Kumar et al. (2018) presents clear guidelines but allows for a broad range of interpretations by the annotators.

Which guidelines are best according to you? Why? Give arguments based on the annotation study.

The guidelines presented in the paper "Predicting the Type of Offensive Posts in Social Media" by Marcos Zampieri et al. (2019) stand out as the most effective due to their detailed explanation of the annotation guidelines. This level of clarity results in more precise annotations and minimises the discrepancies among annotators. This assertion is proven by the high scores obtained in both the Cohen's Kappa test and the Percentage Agreement test, providing compelling evidence of their efficacy.

Appendix:

Contribution:

Topic	Contributor
Sets of annotations	Tanya, Antonios
Inter-annotator agreement scores	Rishikesh
Confusion matrix	Payanshi
Error analysis	Tanya, Antonios, Rishikesh, Payanshi
Comparison annotation models	Tanya, Antonios, Rishikesh, Payanshi