Dataset and Annotation Guideline

For this experiment, we followed the definition proposed by Waseem and Hovy for hate speech and marked the hate span if it consists of any of the following explicit mentions:

- A sexist or racist slur or abusive term attacks a minority group/individual.
- A phrase advocated violent action or hate crime against a group/individual.
- Negatively stereotyping a group/individual with unfounded claims or false criminal accusations.
- Hashtag(s) supporting one or more of the points mentioned earlier.

Additionally, the hate intensity of a sample was marked on a scale of 1-10, 10 being the highest based on:

- Score [8-10]: The sample promotes hate crime and calls for violence against the individual/group.
- Score [6-7]: The sample is mainly composed of sexist/racist terms or portrays a sense of gender/racial superiority on the part of the person sharing the sample.
- Score [4-5]: Mainly consists of offensive hashtags, or most hateful phrases are in the form of offensive hashtags.
- Score [1-3]: The sample uses dark humour or implicit hateful term.

Rephrasing Guidelines:

- Explicit slurs were replaced by “ADJECTIVE + TARGET_GROUP” to reduce the hatefulness which is associated with the slurs due to their extensive use in society over time. Example: “dirty n****** -> dirty black men”. This helps prevent losing the hatefulness
- Strict calls for action such as “Remove XYZ” are replaced by a bit less aggressive messages such as “Scrutinize XYZ.”
- Messages calling out groups asking them to leave and go to place XYZ are made a bit more ambiguous/less hateful by removing target group markers from the place. Example- “If they love Sharia law so much, why do not they go off and live in a Muslim country?” becomes “If they love Sharia law so much, why do not they live in another country?”
- Weaker alternatives like “dislike replace strong words like “hate”.”
- Repetitions of specific hateful phrases are removed to lessen the intensity. Example - “Not BRITISH! Not BRITISH! Not BRITISH!” becomes “Foreigner!”
- Repetitive references to the group are removed to lessen the impact. Example - “Muslims should stay in their country if they want to follow sharia law.” becomes “Some Muslims can stay in their country if they strictly wish to follow religious law.”
- Blanket references are converted to targeted references, targeting some wrong group members and not all. Example - “All crimes that have committed this Muslim gang concern white citizens of England. Only white British are targeted.” becomes “All crimes that some Muslim criminals commit concern citizens of England. Only British are targeted.”
- Blanket labels replace targeted labelling based on crime to reduce intensity. Example - “There are bands of Muslim rapists everywhere in Telford, Rotherham, Rochdale, Oxford, Newcastle etc. It is awful!” becomes “There are Muslim criminals everywhere in Telford, Rotherham, Rochdale, Oxford, Newcastle etc. It is awful!”
- Direct References to Holy Scriptures etc., are made a bit obscure and indirect. Examples - “Thousands of our girls abused by Muslim rape gangs, commanded by Koran. Why do we keep accepting this?” becomes “Thousands of our children are abused by Muslim criminals, commanded by the misinterpretation of their books. Why do we keep accepting this?”
- Probable statements replace conclusive statements. Example - “Muslims are the ones who invented the slave trade.” becomes “Some Muslims might have started forced labour”
- Sometimes there’s no good way to rephrase some hateful parts, so they are just dropped. Example - “Us and UK face challenges from Islam, even if the world is already doomed., These

challenges are cult of death, discrimination, cruelty, liberty, liberty of speech, being gay, Quran, jihad.:" becomes "Us and the UK face lots of challenges from Islam, even if the world is already doomed."

- Targeting groups is replaced by targeting specific members of the group. For example - "Islam is the world's illness and must be cured with antibiotics." this becomes "Some Muslims cause some problems and must be looked into."
- Exaggerating Hashtags are removed. Example - "Muslims are invading London. There are more Muslims than Londoners in London. #londinistan #LondonHasFallen." becomes "Some Muslims are occupying London. There are many of them."

References:

- Waseem and Hovy: https://aclanthology.org/N16-2013/