**Intervention in the user-level interactions (engagements) in a hate speech spread in social networks.**

**RQs:**

1. What do you mean by intervention in the hate speech spread?
2. What are the types of intervention?
3. Can we classify automatically the comments as intervening messages or not (supporting the hate speech)?
4. Can we classify automatically the intervening messages into the various types of interventions defined in (no. 2 above)?
5. What are the types of hate messages (Toxic levels and humiliation kind of hate)?
6. Any relation between the type of hate message and the type of intervention?
7. Can we quantify the hate content before and after intervention so that we can answer if some interventions work or not?

**Objectives of the work:**

1. Define types of interventions.
2. Build a classifier that classifies a comment in a thread as intervening or not.
3. Build a classifier that classifies intervening comments as what type of intervention it is.
4. Using the existing tools classify the original hate post into what type of hate messages.
5. Examine what type of intervention happens for what type of hate post.
6. Quantify the efficiency of various interventions in various types of hate posts.
7. Infer through the quantification method mentioned above what type of intervention works and what doesn’t work.
8. Time series analysis of toxicity before and after interventions.

**Additional one:**

Semantic level analysis of hate speech content using USAS semantic tag or any other equivalent tag sets.