# Hate Speech / Toxic Comment detection

**Group Members:**

1. Harshit Agrawal - 18074019 (CSE IDD Part 3)
2. Ashish Kumar - 18075068 (CSE B.Tech. Part 3)
3. Sachin Srivastava - 18075070 (CSE B.Tech. Part 3)

**Introduction:**

The project aims at improving the user experience of using any website for online chats, conversation and posts by flagging and removing the textual material containing hate and toxicity. These comments are responsible for spreading negativity, harassment of particular races and groups and demeaning people belonging to certain ethnicities. It is therefore the moral duty of the website owner organization to ensure that such comments and posts are taken down as soon as possible in the best interest of everyone and so that everyone can express himself freely over the internet.

**Problem description:**

Given any text or paragraph containing a few lines in natural language (such as English), the objective is to classify it as belonging to one of the following categories:- normal, obscene, threatening, insulting, toxic, severely toxic and hate. As evident this is a multi-class classification problem and from real world knowledge we know that a post can be abusive in multiple ways, for example the same comment may be threatening as well as insulting to someone. Thus this is a multi-label classification problem too. So the model will output the probability of the post belonging to each of the categories. Based on a certain threshold which can be tuned as a hyperparameter, a comment may be classified to be belonging to a category/set of categories.

**Approach:**

As a first look on the dataset, it is easy to spot that there is a significant class imbalance in the dataset. So it is essential to use appropriate metrics to quantify the performance of the model.

We intend to start with certain *visualizations* to get qualitative high level insights into the data. This is essential so that we apply the right algorithms and pattern analysis methodology. Then we will split the dataset into training, validation and testing splits. The training dataset will be used for training the parameters of the model. The validation set is used for hyperparameter tuning and for comparing across the models. Finally the testing set is used only when we have selected the best model using training and validation set. This ensures that we do not overestimate the capabilities of the model. It is further important to shuffle the data well before this splitting so that no bias is introduced in the model due to skewed splitting.

The next important step is *data preprocessing*. Since this is textual data so we need to remove all the punctuation marks which do not contribute significantly to detection of hate speech. Then

some stop words are detected and removed to prevent performance degradation in classifiers based on bayesian methods. We may also remove certain handles and URLs as they don't add any value to the hate speech detection. Finally the words are stemmed and are vectorised so as to enable efficient matrix based processing in machine readable format.

Now we can use a variety of algorithms for *training the model*. We plan to start from very simple algorithms such as the Naive Bayes and decision trees. Based on the results obtained from these models we can move towards application of more complex algorithms such as those based on support vector machines or Recurrent Neural Networks. We also plan to study research papers on related topics to get more ideas about possible algorithms, models and architecture that can suit the problem in hand. We can also try transfer learning approaches and data augmentations if the available dataset seems to be too small to fetch satisfactory results.

After the model is finalised and quantitative results are calculated we can do *post-processing* on the data, that is, we plan to have a look at samples which are wrongly classified by the model and reason the possible errors due to which model went wrong to get an qualitative idea about the performance.

One of the practical problems we are forecasting is limited computation capabilities. Since data mining problems may require significant computational capabilities, the lack of access to institute server and GPU facilities may be an issue for this. Another problem may occur due to limited memory resources of our personal laptops (4-8 GB).

**Applications:**

- The model can flag any rude online behaviour or any unsuitable content on certain social networking websites like Facebook and Twitter, and the respective message or tweet can be automatically reported to the support team and can be removed immediately.

- The model can be used in online meetings, online classes or webinars, so that any toxic or hateful messages or abuses would automatically be hidden from the attendees, and shown only to the moderators or owners of the meeting, who can delete it, if felt unsuitable for the general public.

- The model can automatically flag the contents containing insults to certain race, religion, ethnicity, gender, culture or nationality, and would ensure that the internet is free from these kinds of behaviours.

- The model can help in identifying threatening messages to people or groups of peoples, so that those incidents can be reported to the concerned authorities and dealt with immediately.

- The model can be used in chatbot training by giving large negative feedback if the chatbot generates some reply which may be insulting towards a user.