Toxic Content identification

Azzouz Abd El Djouad Aymen from Group 1.

Introduction:

Social media nowadays has a huge amount of toxicity, so this project has an objective of improving the user experience, by reporting and removing any comments or messages that contains any particular form of hate. Belonging to any race is in no way disrespectful and no one should be bullied or harassed for it. Providing solutions for this problem is every organization's responsibility, that such things cannot be accepted in any way or form, for the sake of freedom, and freedom of speech.

Problem description:

The job is to take any given Text or Paragraph, that contains different lines in NL (we have English language for this) and classify it, it could be normal or hate, toxic or severely toxic, obscene or threating, maybe insulting. As we can see this is a multi-class classification type of problem, using logic, a comment could be in more than just one class, threat is usually insulting as an example, so it is a problem of multi-label classification too. Using a measure that could be tweaked to classify a comment in a category or a set of different categories.

Approach:

Taking a first look into the dataset, class imbalance could be spotted significantly in the dataset. So, it is a must to quantify the performance of the model using appropriate ways and metrics.

Starting with some visualization, to give us an accurate look into the data. Knowing this will help us a lot to apply the right algorithms. Then we'll create a training set, a test set, from the original dataset

Training the parameters of the model using the training set. Comparing, across the models using the 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.

Data preprocessing is an essential step that cannot be skipped. Since this is textual data, punctuation points are no use for us, so we will be removing them. Detecting stop words and removing them to avoid performance deficiency in classifiers based on Bayesian methods. Handles and URLs will be removed too, no use for us again. Finally, stemming words and vectorizing them, to enable efficient matrix-based processing. After that, algorithms are ready to be applied on the training data set. We'll start with the Naive Bayes. Depending on the results from this model, we'll apply a more complex algorithm.

Applications:

The model can:

● flag any online behavior that is found unsuitable content on certain social media, like Instagram, and the respective comment/message can be reported, automatically, to the support team for an immediate action.

● be used in online meetings/classes or webinars, so that any hateful content or abuse, would be hidden, automatically, from the attendees. shown only to the moderators.

● automatically flag the contents containing discrimination or racism to certain race/religion/ ethnicity/gender/culture or nationality, ensuring the safety of mental health of everyone.