**Toxic Comments Detection and Classification:**

**Case Study of Wikipedia Discussions**

**Introduction**

In most online communities, social interaction and discussion are core mechanisms, through which users interact with each other and share information and opinions on some topics. Social interactions happened not only on social networks such as Facebook or Twitter, they also are key mechanism in crowdsourcing and peer production communities. For example, in Wikipedia, volunteer editors create knowledge collectively, and it is through social interaction that many editors could coordinate and collaborate with each other to generate plans, to discuss with each other, to provide useful feedbacks, and to eventually create a featured article.

However, the anonymity afforded by such online communities has led to increase of misbehavior, such as abuse and harassment, spread of propaganda, hate speech, and many more. In discussion oriented social media platform, the threat of abuse and harassment online means that many people stop expressing themselves and give up on seeking different opinions. In peer production communities such as Wikipedia, such toxic behaviors can even impact the volunteers’ willingness to future contribution: over half of the victims reported decreased participation after the harassment[[1]](#footnote-1). Previous research also suggested that being exposed to harassment tend to bring negative emotions to the community[[2]](#footnote-2).

In order to facilitate discussions and conversations with in the online communities and to maintain a more healthy online environment, current project aims to study the toxic online comments (i.e. comments that are rude, disrespectful or otherwise likely to make someone leave a discussion). Specifically, the current project intend to detect the toxic social interactions and classify the toxic social interaction into different types of toxicity.

**Dataset**

I am going to use the dataset from Kaggle competition: <https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge>. This data includes a large number of Wikipedia talk page discussions which have been labeled by human raters for toxic behavior. There are six types of toxicity: toxic, severe toxic, obscene, threat, insult, identity hate. The current project aims to create models which can both detect toxic comments from ordinary comments, and predict a probability of each type of toxicity for each comment.

**Method**

I plan to build two set of model.

* Binary classification: The first set of model is to simply detect whether the current comment is a toxic comment or not, and this model is a binary classification model.
* Multiclass classification: The second set of model is than able to do multi-classification task by providing probability score for each type of toxicity.

The specific steps I will take include:

1. Data cleaning and exploration
2. Feature extraction for **basic features set**, including length of the comment, number of capitals, number of exclamation marks, question marks, punctuation, number of symbols (more symbols in foul language (e.g., f\*ck or $#\* or sh\*t)), number of words, number of I, You, We, She, He., etc.
3. Feature expansion for **expanded feature set** using additional dictionary and word embedding
   1. LWIC dictionary with different emotional lexicons
   2. Trained word embedding from NLP, Stanford: https://nlp.stanford.edu/projects/glove/

For both model, I plan to use a **simple logistic regression with basic feature set** for baseline model. I plan to furfure use **expanded feature sets** test with other classification methods including SVM, Generative models, decision trees, and neural network, and then use **ensemble methods** to bundle different classification methods. In addition, I plan to build a **deep neural network** model using Tensorflow to test for model performance. I will use accuracy rate to evaluate the models.

**Timeline:**

03/07 – 03/20: Data cleaning, basic feature set, expanded feature set will be ready

03/21 – 04/01: Binary Classification models ready

04/01 – 04/20: Multiclass classification models ready

04/21 – 04/26: Model improvement and final preparation for submission

1. Support and Safety Team. "Harassment survey", [*Wikimedia Foundation*](https://meta.wikimedia.org/wiki/Wikimedia_Foundation), 2015. [↑](#footnote-ref-1)
2. https://meta.wikimedia.org/wiki/Research:Study\_of\_harassment\_and\_its\_impact#cite\_note-1 [↑](#footnote-ref-2)