# Project Proposal

**Project Title:** Fight Online Abuse Using Natural Language Processing

**Team Members:** Aimen Chaudhry and Kalyan Bandaru

***Goals and Objectives:*** The project goal is to detect the type of comment toxicity during online interaction. Given the rise in trolling and hatred on social media and elsewhere these days, this is a very real problem. The task is set up as a standard text classification problem in which one have to predict the likelihood of a (text) comment being toxic or a threat, and submissions are evaluated. The project will be completed by the team to allow for effective brainstorming on how we can improve the model accuracy and how we can lend the NLP to be able to fight online abuse that may have a negative impact.

**Significance:** In today's world, social media has become an integral part of life. Every day people are not only interacting with many people including strangers. This is where things get a little worrisome. With the increasing number of people on social media, there is also an increase in the number of people who do criminal activities like cyberbullying, harassment, etc. The threat of online abuse may result people to stop expressing themselves and seek others comments. Platforms struggle to effectively facilitate conversations, leading many communities to limit or completely shut down user comments. Data reveal that online abuse is a problem mostly affecting Teens by things that are posted by other people either the people they know or strangers.

**Features:** Detecting toxicity is a lot more than just detecting abusive words in the text. As an example, consider the following comment, which doesn't have any abusive words, but is still toxic the approach used is supervised since we have the labeled data.

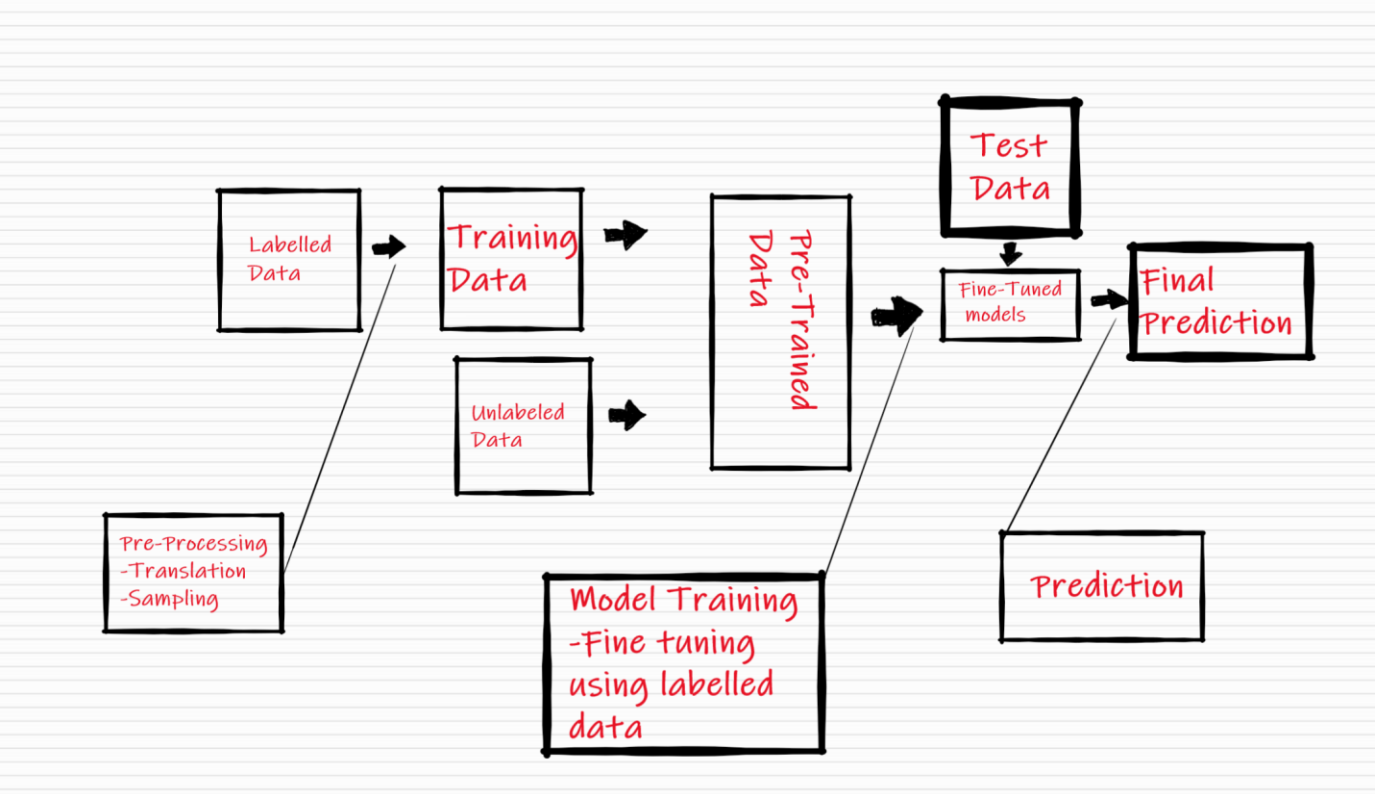


**Fig. 1. Showing how Data will be handled.**

## REFERENCES

[1] *Guynn, Jessica (16 February 2016). "Google Ideas to become Jigsaw tech incubator". USA Today. Archived from the original on 16 August 2020. Retrieved 14 February 2017.*

[2] Canadian Centre for Child Protection. (2016, January). Child Sexual Abuse Images on the Internet: A Cybertip.ca Analysis. p. 19. Retrieved from: https://www.protectchildren.ca/pdfs/CTIP_CSAResearchReport_2016_en.pdf

[3] Retrieved from: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

[4] Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In Proceedings of the International Conference on New Methods in Language Processing, pages 44–49.

[4] Maxim, D., Orlando, S., Skinner, K., & Broadhurst, R. (2016). Online Child Exploitation Material–Trends and Emerging Issues: Research Report of the Australian National University Cybercrime Observatory with the Office of the Children's Safety Commissioner

[5] Retrieved from https://www.analyticsvidhya.com/blog/2019/07/how-get-started-nlp-6-unique-ways-perform-tokenization/

[6] - J. Zhang, Y. Li, J. Tian, and T. Li, "LSTM-CNN Hybrid Model for Text Classification," 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, 2018, pp. 1675–1680, doi: 10.1109/IAEAC.2018.8577620.

[7] Pelfrey, W.V., Jr.; Weber, N.L. Student and School Staff Strategies to Combat Cyberbullying in an Urban Student Population.Prev. SchFail:Altern.Educ.Child.Youth2015,59,227–236, doi:10.1080/1045988X.2014.924087.

[8] Smets, K.; Goethals, B.; Verdonk, B. Automatic vandalism detection in Wikipedia: Towards a machine learning approach. In Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy; Chicago, USA; AAAI Press; 2008; pp. 43–48.

[9] Retrieved from https://towardsdatascience.com/a-step-by-step-guide-to-completely-learn-data-science-by-doing-projects-d7b6a99381ef.