# Introduction

### Literature Review

#### Regression approaches for sentiment analysis

Logistic regression is widely used on natural language processing and has a strong support of theory and algorithm. For example, in topic classification, text categorization, detection of spam pages[3]. Logistic regression measures the relationship between an output variable Y(category) and one or more independent variables, which are usually continuous, by using probability scores as the predicted values of the dependent variable.

Recent studies show that for a wide range of text classification tesks and corpora, logistic regression achieves outstanding performance. In Zhang and Oles 2001 paper[2], they applied logistic regression on text categorization problem and compared with other linear classification methods such as support vector machines, which showed an excellent performance on most of the experiments. In Li and Yang 2003 paper[1], they optimized loss function and model complexity for a set of algorithm include logistic regression, and applied modified approches on text categoritzation problem. The performance of logistic regression is competitive to many latest techniques.

Many efforts also have been made on speeding up training process. A large-scale Bayesian logistic regression approach was proposed on 2007 by Alexander and Lewis[9]. 4-gram hashed byte are used as features to train logistic regression classifier mainly on news articles and showed a great success on text classification problem of large dataset. In 2012, Ke and Meng, et al proposed LightGBM[8], a highly efficient gradient boosting decision tree algorithm, which can achieve an even better training speed on logistic regression with large dataset.

These new techniques allow us to solve the problem of toxic comment classification and identification by using logistic regression.

# Methedologies

### Gradient Boosting Decision Tree Classifier

Gradient boosting decision tree(GBDT)[4] is an optimized distributed gradient boosting framework. Because of its high efficiency and accuracy, it becomes more and more popular for solving various machine learning problems, for example, ranking problem[7], user behavior prediction[6] and multi-class classification[5]. There are several implementation of GBDT recently, such as XGBoost[10], pGBRT[] and LightGBM. In this project, we used LightGBM to implement our logistic regression model.

To further improve the computational efficiency, lightGBM applied Gradient-based One-Side Sampling(GOSS) and Exclusive Feature Bundling(EFB) strategy to exclude non-informative data instances and features so that improve the training speed greatly without losing accuracy. GOSS keeps data instances with larger gradients and randomly drop instances with small gradients to retain the accuracy of information gain estimation and speed up simultaneously. EFB are designed to dealing with datasets that have a large but sparse feature space. By greedily bundling mutually exclusive features, EFB can reduce the number of features.

Logistic regression is widely used on natural language processing, for example, in topic classification, text categorization, detection of spam pages. Logistic regression measures the relationship between an output variable and one or more independent variables, which are usually continuous, by using probability scores as the predicted values of the dependent variable. For input feature vectors , outcome ranges between 0 and 1 which represents the prediction of probability. For probability greater than 0.5, we would classify the observation as positive, otherwise, as negative. The probability of the out is assumed to follow a parametric model given by the sigmoid function:

Reference:

[1] Li, Fan, and Yiming Yang. "A loss function analysis for classification methods in text categorization." *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. 2003.

[2] Zhang, Tong, and Frank J. Oles. "Text categorization based on regularized linear classification methods." *Information retrieval* 4.1 (2001): 5-31.

[3] Genkin, Alexander, David D. Lewis, and David Madigan. "Sparse logistic regression for text categorization." *DIMACS Working Group on Monitoring Message Streams Project Report* (2005).

[4] Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." *Annals of statistics* (2001): 1189-1232.

[5] Li, Ping. "Robust logitboost and adaptive base class (abc) logitboost." *arXiv preprint* *arXiv*:1203.3491 (2012).

[6] Richardson, Matthew, Ewa Dominowska, and Robert Ragno. "*Predicting clicks: estimating the click-through rate for new ads." Proceedings of the 16th international conference on World Wide Web.* ACM, 2007.

[7] Burges, Christopher JC. "From ranknet to lambdarank to lambdamart: An overview." *Learning* 11.23-581 (2010): 81.

[8] Ke, Guolin, et al. "LightGBM: A highly efficient gradient boosting decision tree." *Advances in Neural Information Processing Systems.* 2017.

[9] Genkin, Alexander, David D. Lewis, and David Madigan. "Large-scale Bayesian logistic regression for text categorization." *Technometrics* 49.3 (2007): 291-304.

[10] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* ACM, 2016.

[11] Tyree, Stephen, et al. "Parallel boosted regression trees for web search ranking." *Proceedings of the 20th international conference on World wide web*. ACM, 2011.