# Sentiment classification and usefulness prediction of Yelp reviews

**Authors:** Aditya Priyadarshi, Aditya Sridhar, Yogesh Gupta and Shubham Sharma

**Problem Description:** The objective of this project is to train a classifier that classifies the user given reviews in natural language into positive or negative categories for various restaurants in the Yelp dataset [1]. In addition, we are also planning to undertake the task to determine whether any particular user review can be categorized as useful or not.

**Summary of Data:** The original dataset described in the Yelp Dataset Challenge 9 [1] has 4.1M reviews and 947K tips by 1M users for 144K businesses spread across four cities. We are planning to use a subset of the given data. In our initial experiments, we have taken first 5000 reviews. The anonymized data is provided in JSON format for each review and each user profile. In the initial set of 5000, we had no missing data. Each individual review data consists of anonymized IDs for the business, user and review, star rating, review type, review text and votes on how useful, funny or cool the review is. The relevant variables in the user data consists of review count, the number of fans, the number of useful, funny, cool votes the users reviews have received in addition to other attributes.

```
{
    "review_id":"encrypted review id",
    "user_id":"encrypted user id",
    "business_id":"encrypted business id",
    "stars":star rating, rounded to half-stars,
    "date":"date formatted like 2009-12-19",
    "text":"review text",
    "useful":number of useful votes received,
    "funny":number of funny votes received,
    "cool": number of cool review votes received,
    "type": "review"
}
```

Figure 1: Review Data Format

**Methods:** We plan to use Naive Bayes and SVM to train our classifier as discussed in paper [2] using bag-of-words feature. In addition, we also intent to use convolutional neural networks (CNN) and recurrent neural networks (RNN) as discussed in papers [3] and [4] respectively using word2vec feature.

**Preliminary results:** In our initial experiment, we tried Naive Bayes classifier for sentiment classification using Python. We took first 5000 reviews in the dataset and divided into training set (80%) and validation set (20%). Reviews with star rating of 3,4 and 5 were considered positive and 1,2 as negative. We have then selected 6790 words from selected corpus as bag-of-words features which have term frequency of five or more and document frequency of two or more. Naive Bayes implementation provided by NLTK was used for training classifier. On the validation set, we have received an accuracy of 77.8% using Naive Bayes for sentiment classification into positive and negative reviews.

## References

[1] Yelp Dataset Challenge `https://www.yelp.com/dataset_challenge`

[2] Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. *Thumbs up? Sentiment Classification using Machine Learning Techniques*, In EMNLP, 2002.

[3] Yoon Kim. *Convolutional Neural Networks for Sentence Classification*, In EMNLP, 2014.

[4] Dani Yogatama, Chris Dyer, Wang Ling, Phil Blunsom. *Generative and Discriminative Text Classification with Recurrent Neural Networks*, arXiv preprint arXiv:1703.01898, 2017