Novel usefulness classification system for online reviews

Speech and Natural Language Processing
Term Project

Project ID:2
Khadpe Pranav Ajay(15EE30025)
Atul(15EC10067)
Joshua Peter Ebenezer(15EC10023)

Mentor: Soumya Sarkar

The problem

Reviews written in sites such as yelp.com' often get no reads at all due to the huge number of other reviews, regardless of whether they are actually useful or not.

Users can give a 'useful', 'funny' or 'cool' vote to every review on yelp. We thought it would be great if we could actually predict whether a review would be rated 'useful' by a lot of people, and if we could then use that to float reviews that are very likely to be useful to others to the top.

The data

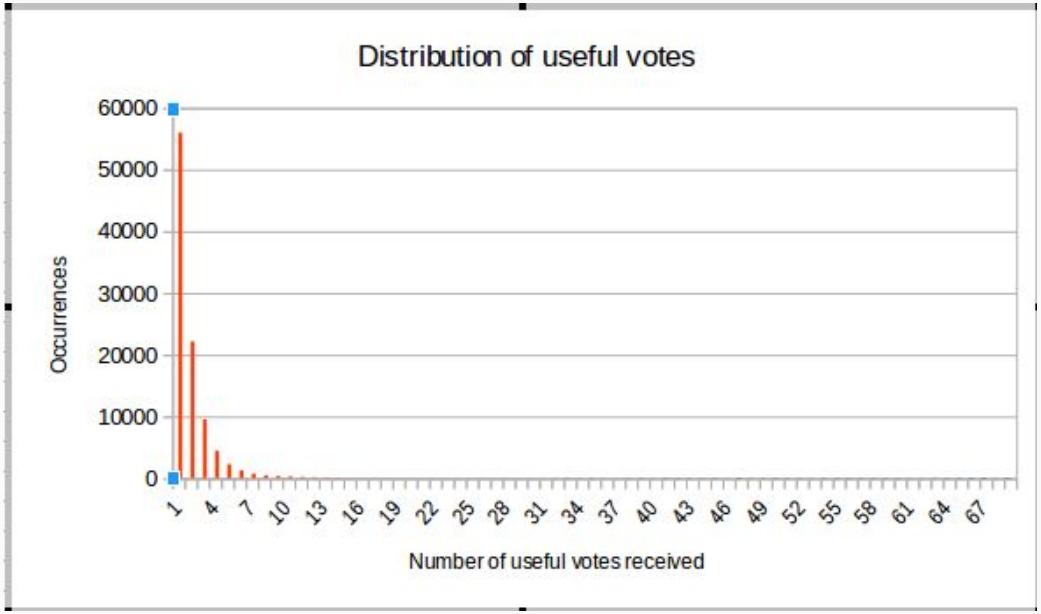
Yelp released a raw dataset of around half a million reviews to the public in September 2017 for it’s 10th dataset challenge, and invited submissions for analysing the data and arriving at useful conclusions or applications.

The text of each review, the user id, the number of useful, funny and cool votes received, the business id, the rating, and the time the review was written were made available.

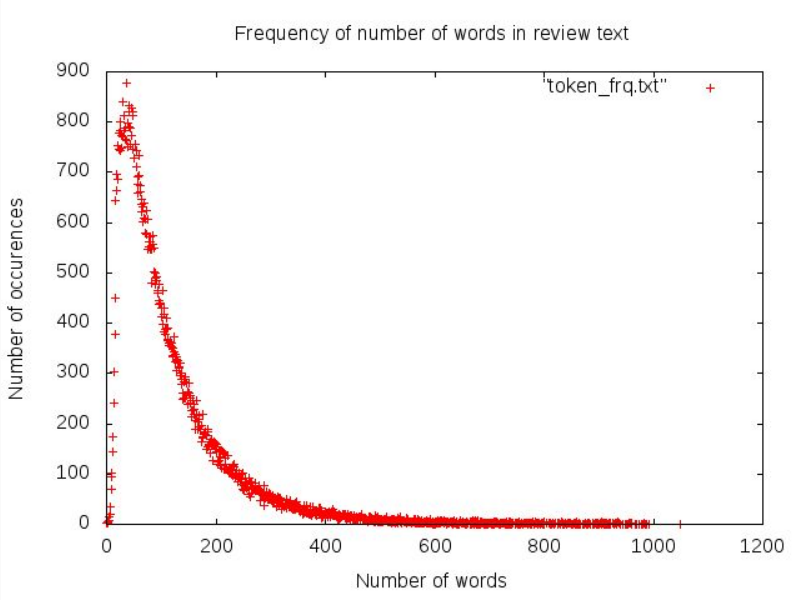
Looking at the data

- 45.2% reviews have 0 upvotes and 98.6% have less than 20.
- A review voted to be cool or funny may not be useful but they ensure that the review has been seen.
- Maximum number of reviews contain about 35-40 words.
- Most reviews have 0 useful votes - might be because they were not viewed or they were genuinely not useful.

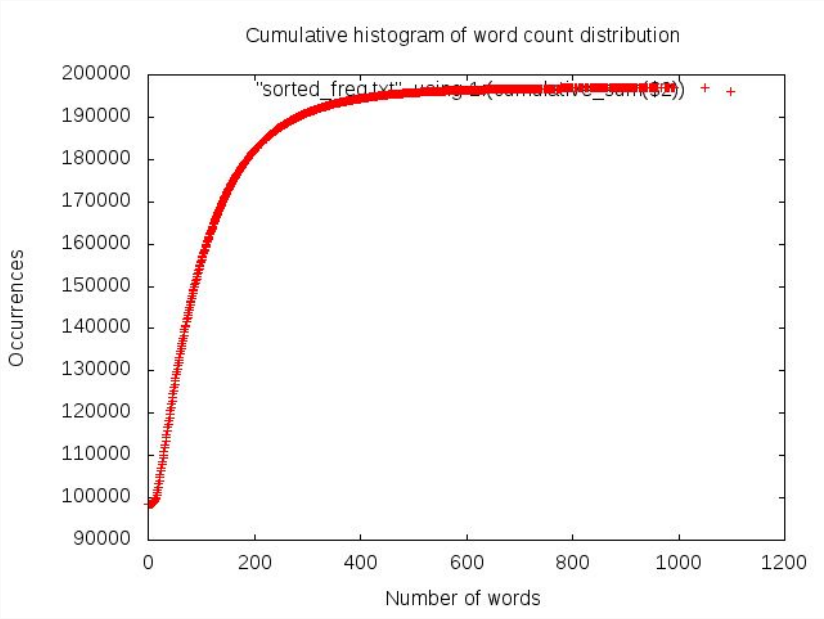
Frequency distribution of useful votes



Distribution of the wordcount of reviews



Cumulative histogram of wordcount



Preliminary results - softmax with tfidf

We randomly sampled 100,000 reviews and binned them into 2 levels (of useful votes received).

We found tf-idf features, and performed 5 fold cross validation with the softmax model.

We achieved 56% accuracy.

Eliminating the effect of time

New reviews may not have been read even if they were useful

We removed all reviews written after January 1st 2017 and before January 1st 2015 to remove old reviews (that may no longer reflect recent trends) and new reviews (that may not have been seen by more than a few people).

Eliminating reviews that were never seen

We kept only those reviews that have upvotes in either funny, cool, or useful. This ensures that all the review was seen, but still allows non-useful reviews.

Eliminating reviews that are tiny or huge

Kept reviews having word count between 10 and 500.

Eliminating non english reviews

Removed non English reviews using the langid classifier.

Two approaches:

Multi class classification after labelling into 4 bins.

Binary classification after labelling into 2 bins

Multi Class classification

Dataset: 5920 reviews equally distributed among 4 bins.

Features: LIWC Features

Features reduced using ExtraTrees classifier

Result: Average Accuracy 40.5%(Random Forest)

Breakdown of result: Test size 1184

Correctly predicted:480

Off by 1: 440

Off by 2:190

Off by 3: 74

Binary Classification

Dataset: 5920 reviews into 2 bins

Feature: LIWC features

Result: Average accuracy: 75%

Best accuracy: Logistic regression (76%)

Interpreting the results

In the multiclass approach the classes are not clearly demarcated. In the sense a classification off by one is still giving us some information about the review and can not be neglected as incorrect.

The separation between 2 bins is often just one vote.

Interpreting the results

Another way to interpret the results is to use the binary classifier (which gives 76% accuracy) and quote the probability of the review being classified as ‘useful’ as a measure of usefulness, which is also reasonable.