# Feature Extraction System

**Project Proposal:**

**Extracting the features that affect the reviews and ratings on yelp negatively. ( focussed on restaurants )**

Sravan Kumar Reddy Mummadi

**Objective**

Model a system that extract features based on user reviews for businesses that affects ratings and reviews of businesses. (focussed on restaurants)

**Abstract:**

Yelp allows users to rate on parameters like service,food .. etc., Often a rating doesn't help much in understanding the positives and negatives of a restaurant. My system helps in summarizing a review and extracting the features of a business. This helps for business owners to understand the factors that affect their rating and reviews on yelp. I used nltk POS tagger to parse the review and wrote regex grammar rules to extract specific patterns of a review and check with negative words corpus, that I have prepared by manually reviewing 200 negative reviews.

**Project components:**

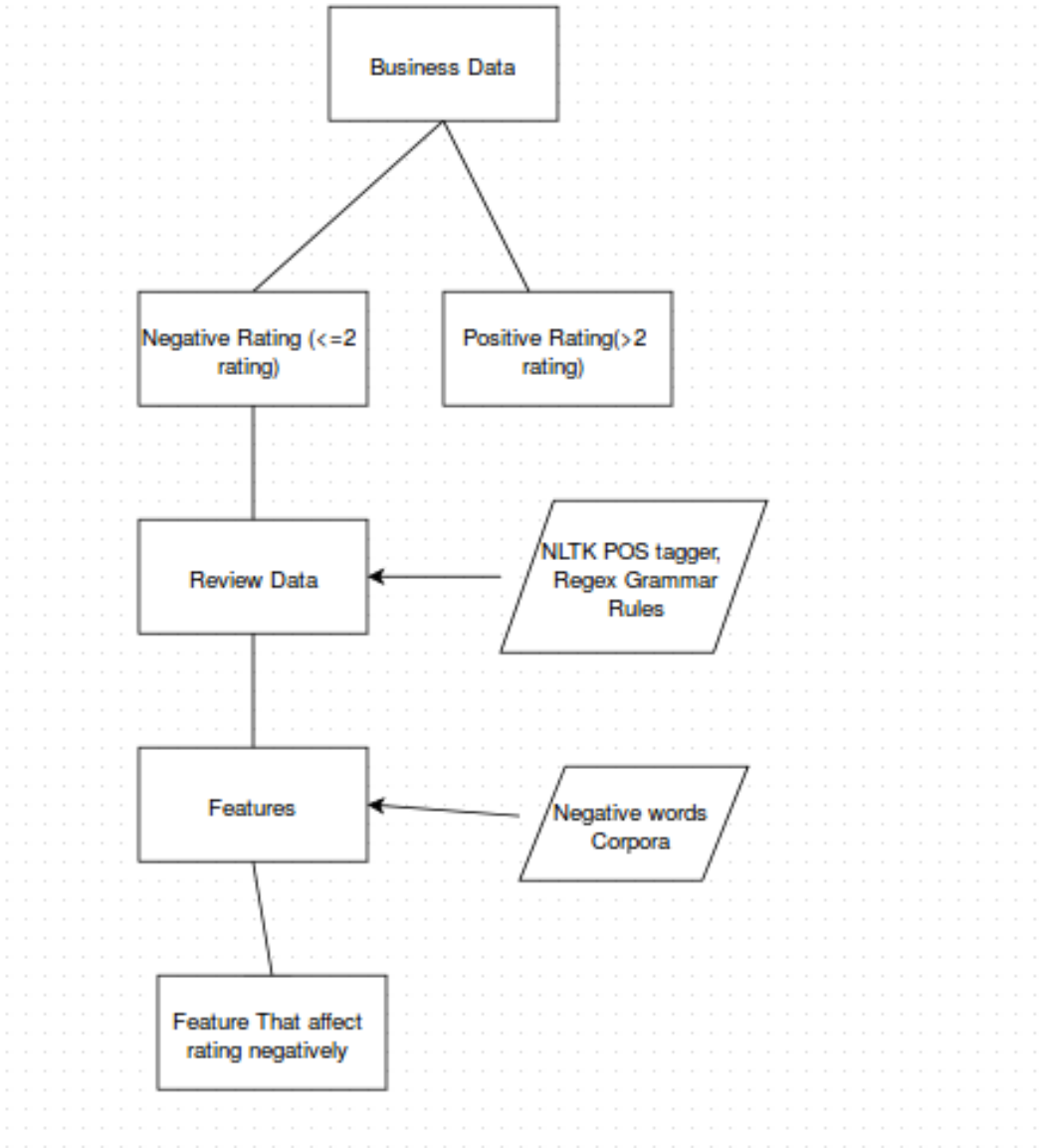
**Negative words corpora:** I have prepared negative word list corpora manually by looking at 200 reviews for the businesses which has poor rating.

**negative_words_corpus=**['not good','bad','very small','tough','expensive','raw','mediocre','dirty','horrible','filthy','awful','gross','unpleasant','dead','greasy','average','nasty','so-so','dry','meh','hard','high','ridiculous','too much','poor','sub-par','crappy','wrong','undrinkable','yucky','sketchy','enormous','rubbery','not impressive','far','not edible','not close', 'offensive', 'loud', 'unreasonable', 'slow', 'tough', 'weird','heavy','overly','messy','incorrect','non-existent','inappropriate','unprofessional','opposite','terrible','wasn't fresh','not fresh', 'rude', 'huge','tiny', 'lousy', 'uncomfortable','inconvenient','busy','sad']

**Files:**

FeatureExtraction.py: It parses all the reviews and extract the features.
BusinessData_GoodBadRating.py: It creates separate files of good and bad rating for restaurant business.

**Flow Chart:**


Business Data
Negative Rating (<=2 rating)
Positive Rating(>2 rating)
Review Data
NLTK POS tagger, Regex Grammar Rules
Features
Negative words Corpora
Feature That affect rating negatively

**Regex Patterns:**

patterns = """attribute_review: {<NN><VBD><JJ>}
{<NN><VBZ><JJ>}
{<VBD><DT><VBZ>}
{<NN><VBD><RB>*<JJ>}
{<RB><JJ><NN>}"""

**Results:**
**Features  Extracted for each restaurant:**

'l2gPB9mqiHSbpwSUa7zrjg': [' d was reluctant', ' charge was deliberate', " food was n't as good", ' food is great', ' bread was average', ' service was average', ' practically next door', ' waitress was nice', ' rib was fatty', ' fee is exorbitant', ' bread was very sorry', ' greeter was very nice', ' server was very knowledgeable', ' extremely strict diet', ' very good (', ' steak was good', ' service was slow', ' zucchini is great'],

'bZcqORBnVApUA2-SEn7VEQ': [' food is consistent', ' wifi was slow'],

'BbIh5NTizhV4Fq_mLmNkpg': [' place is ridiculous', ' sparingly frequent fast', ' very long time'],
'e2K0YQel5Fth0_vur2dN8w': [' here last year', ' food is good', ' place was so bad'],
'xUf11yTcoRagwNiJcY8GAA': [' something is outstanding', ' food was typical', ' service was horrible'],

'PK6aSizckHFWk8i0oxt5DA': [' service is terrible', ' food is hot', ' manager was rude'],

'5vLVIomtminS_q8itJquEQ': [' food is awful', ' generally nasty )', ' service is slow', ' place has such', ' place is filthy', ' food is atrocious', ' everything is fresh'],

'qJTHaHFKQIKXKX3I17Nw9w': [' probably terrible management'],

'GSiHJG8LqTn5ZQAY1r9q9w': [' buffet was really terrible', ' service is good', ' pretty bad chinese', ' buffet is bad'],

'EZrCQtZxiEo1kkAYt2EQqw': [' location is horrible', ' order is wrong'],

'6ilJq_05xRgek_8qUp36-g': [' food is good', ' expectation is quick', ' mostly tame partying', ' milkshake was runny', ' milkshake was good', ' morbidly obese hostess', ' server was high', ' cleanliness is fair', ' sometimes outright rude', ' food was actually pretty good', ' food is good', ' food is good', ' food is good', ' food is good', ' service is awful', ' service was fast', ' food was

terrible', ' burger was really greasy', ' food was finally ready', ' cashier was rude', ' alone is worth', ' location is terrible', ' burger was not fresh', ' really good looking', ' really clean either']}

PK6aSizckHFWk8i0oxt5DA [' service is terrible', ' manager was rude']
GSiHJG8LqTn5ZQAY1r9q9w [' buffet was really terrible', ' pretty bad chinese', ' buffet is bad']
4LcFKTr6Ah87VcxW2z5e6w []
BbIh5NTizhV4Fq_mLmNkpg [' place is ridiculous']
xUf11yTcoRagwNiJcY8GAA [' service was horrible']
6ilJq_05xRgek_8qUp36-g [' server was high', ' sometimes outright rude', ' service is awful', ' food was terrible', ' burger was really greasy', ' cashier was rude', ' location is terrible']

l2gPB9mqiHSbpwSUa7zrjg [' bread was average', ' service was average', ' service was slow']
bZcqORBnVApUA2-SEn7VEQ [' wifi was slow']
EZrCQtZxiEo1kkAYt2EQqw [' location is horrible', ' order is wrong']

**Conclusion:**

I have successfully extracted features from all the review set using the nltk regex patterns and successfully extracted the negative features using the corpora.