

---

# **Yelp Review Rating Prediction**

Deepika, Jinlong, Connor, Lin

---

# Business Applications



- According to Inc. Magazine, 84 percent of people trust online reviews as much as friends' recommendations when it comes to making purchasing decisions.
- Every consumer-facing business has an online presence that can be impacted by customer reviews.
- Businesses must actively manage their online reputations in order to succeed in the market.
- By building models that predict the review rating and sentiment score, businesses will be better able to manage customer complaints by intelligently allocating customer service resources.
- Our methods will reduce the burden of negative reviews by flagging customers who are likely to complain online.

---

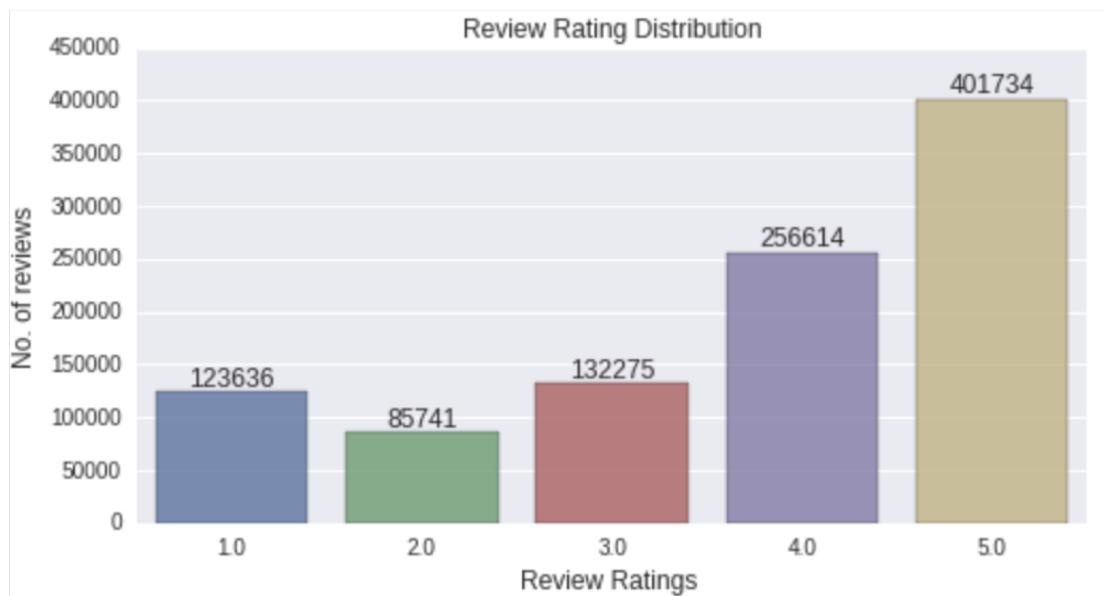
# Data Overview

- ❖ Dataset taken from <https://www.kaggle.com/yelp-dataset/yelp-dataset>.
- ❖ Tables like business, review, user, check in, tip existed.
- ❖ Combined business, review and user tables containing records for users, their reviews and the information about the businesses users have reviewed.
- ❖ 1 million rows having columns like review id, user id, business, review, city, etc. were considered.

---

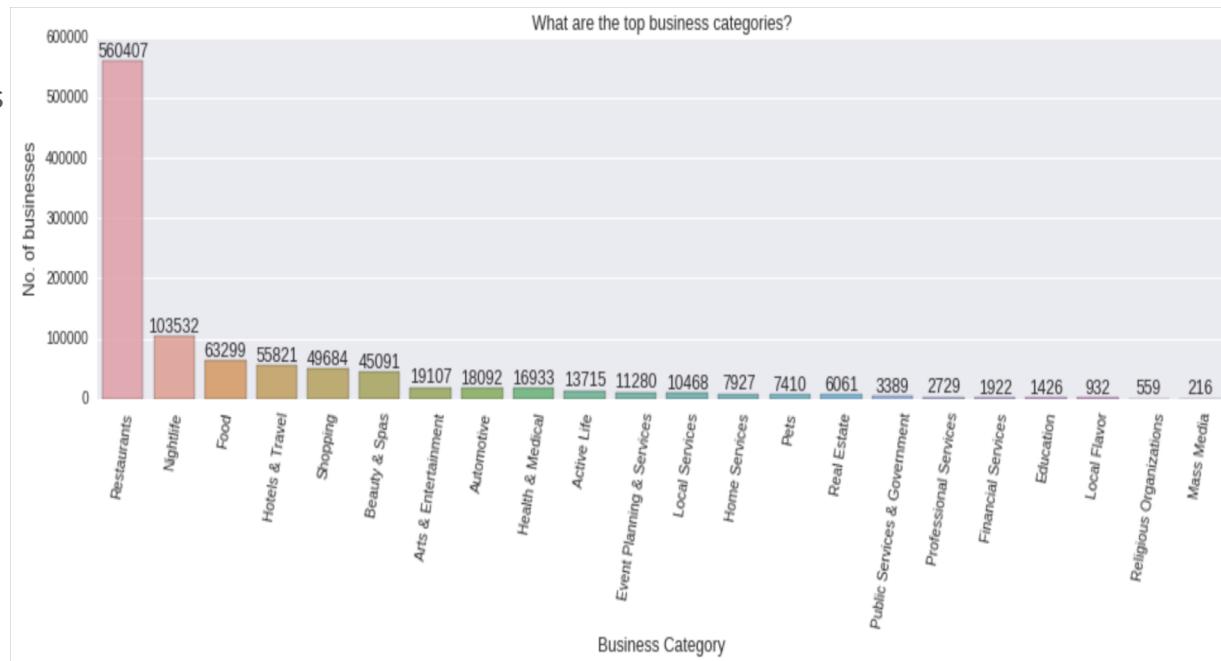
# Exploratory Data Analysis

- No. of reviews by rating



# Exploratory Data Analysis

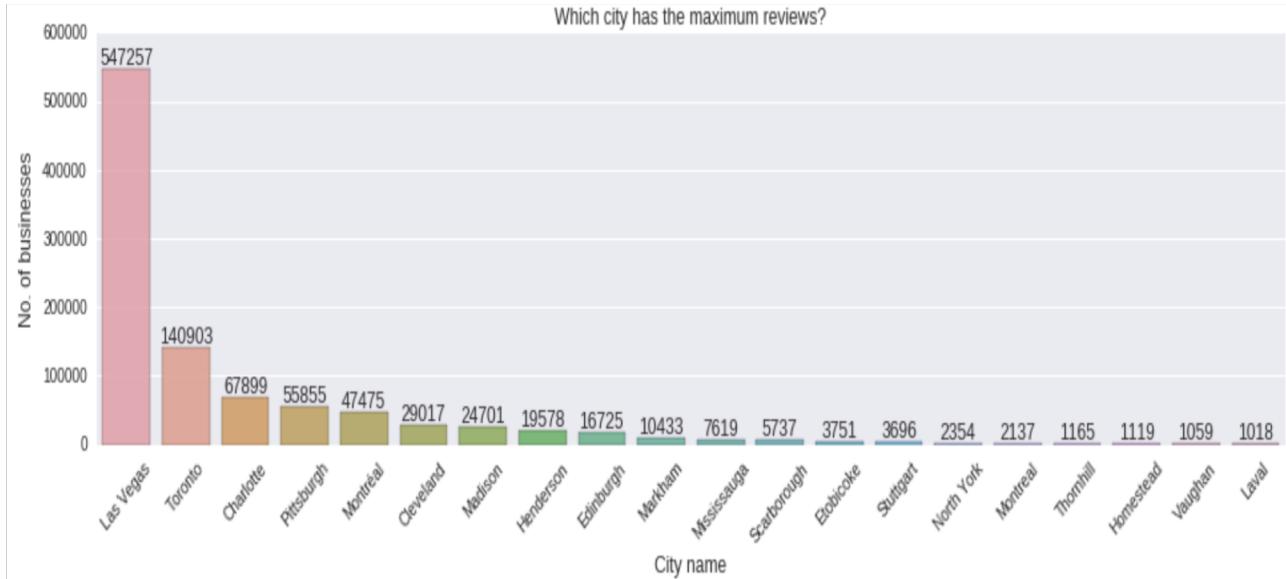
## □ Popular business categories



---

# Exploratory Data Analysis

## No. of reviews by cities



# Exploratory Data Analysis

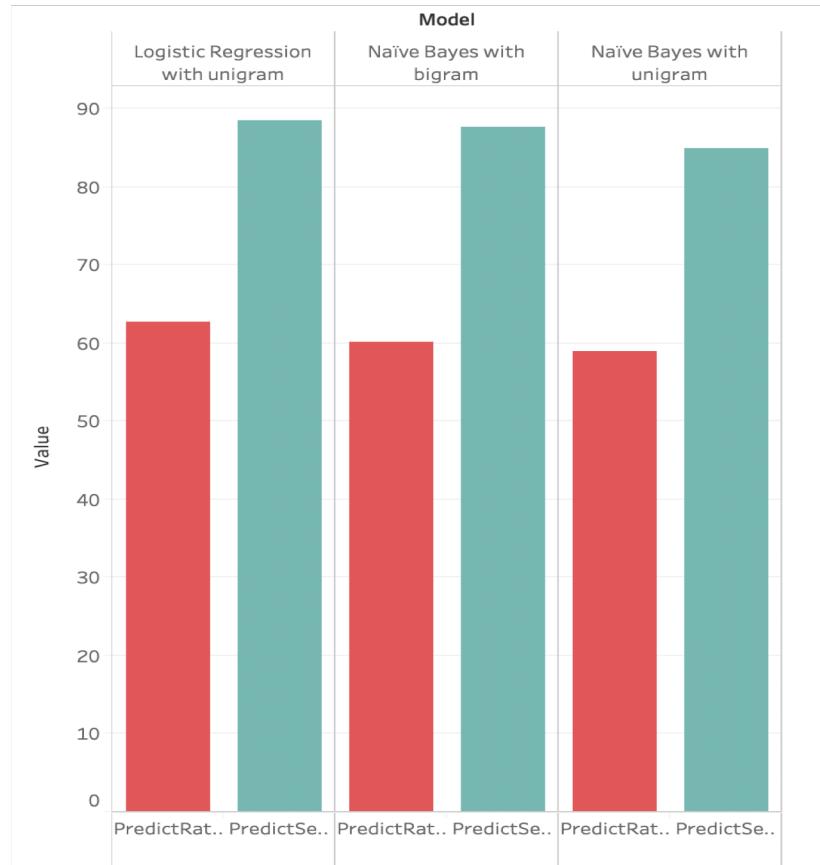
- Most common words in reviews





# N-Gram

Model	PredictRating(%)	PredictSentiment(%)
Naïve Bayes with unigram	58.95	84.86
Logistic Regression with unigram	62.65	88.52
Naïve Bayes with bigram	60.16	87.58



# LDA Model Introduction

**Topics**

- gene 0.04
- dna 0.02
- genetic 0.01
- ...

- life 0.02
- evolve 0.01
- organism 0.01
- ...

- brain 0.04
- neuron 0.02
- nerve 0.01
- ...

- data 0.02
- number 0.02
- computer 0.01
- ...

**Documents**

**Topic proportions and assignments**

## Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game; particularly as more and more genomes are being fully mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains

Araday Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing all

\* Genome Mapping and Sequencing. Cold Spring Harbor, New York, May 8 to 12.

The diagram illustrates the process of determining the minimum number of genes required for life. It shows four stages:

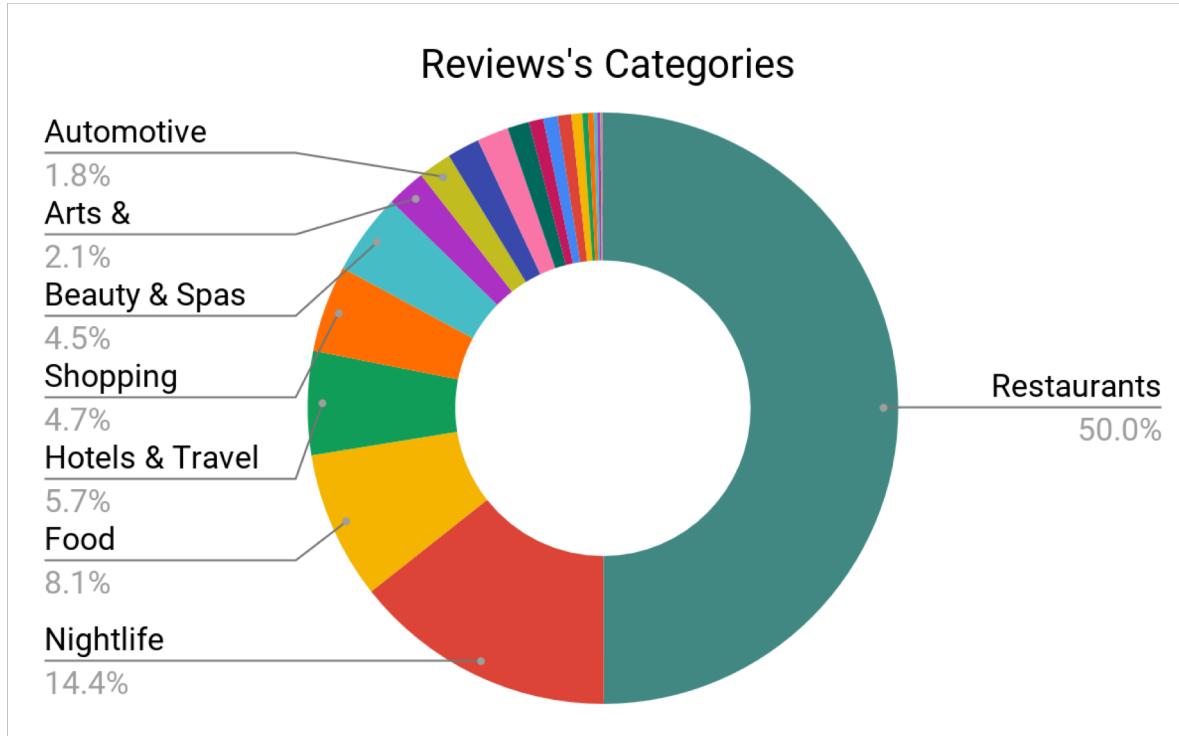
- Haemophilus genome (1700 genes)
- Genes needed for basic life (~220 genes)
- Redundant and parasite species (~22 genes)
- Basic life genes (~4 genes)
- Minimal gene set (250 genes)

Stripping down computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

# Reviews segmentation based on categories

---



# LDA Topics

---



“ Super simple place but amazing nonetheless. It's been around since the 30's and they still serve the same thing they started with: a bologna and salami sandwich with mustard. Staff was very helpful and friendly.”

**63%**

Place, Food,  
Good, Great,  
Service, Love,  
Really, Price,  
Best, Delicious

**17%**

Chicken, Fry,  
Burger, Sauce,  
Sushi, Rice,  
Roll, Soup,  
Order, Dish

**14%**

Pizza, Cheese,  
Dessert, Steak,  
Bread, Salad,  
Sauce, Dish,  
Wine, Potato

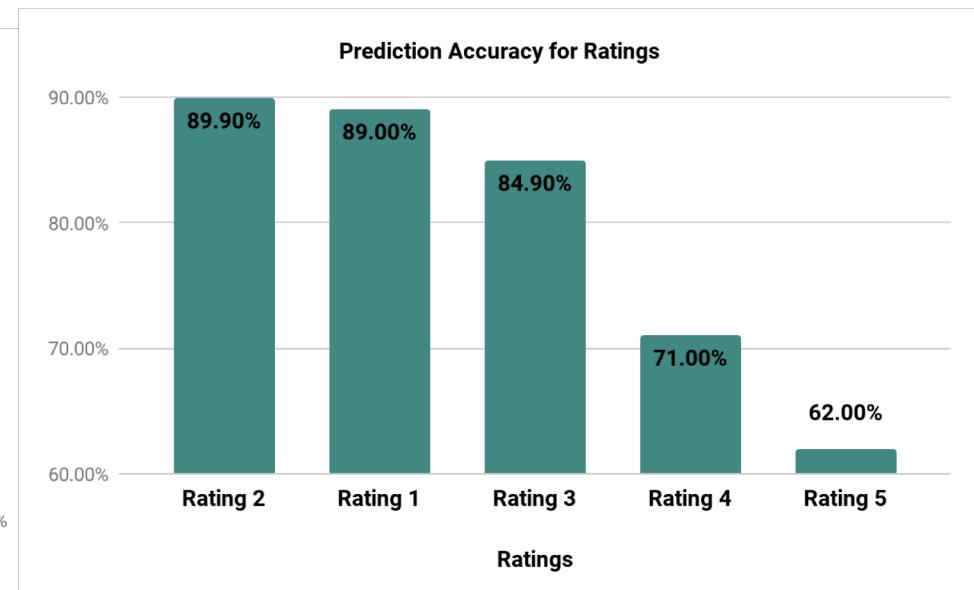
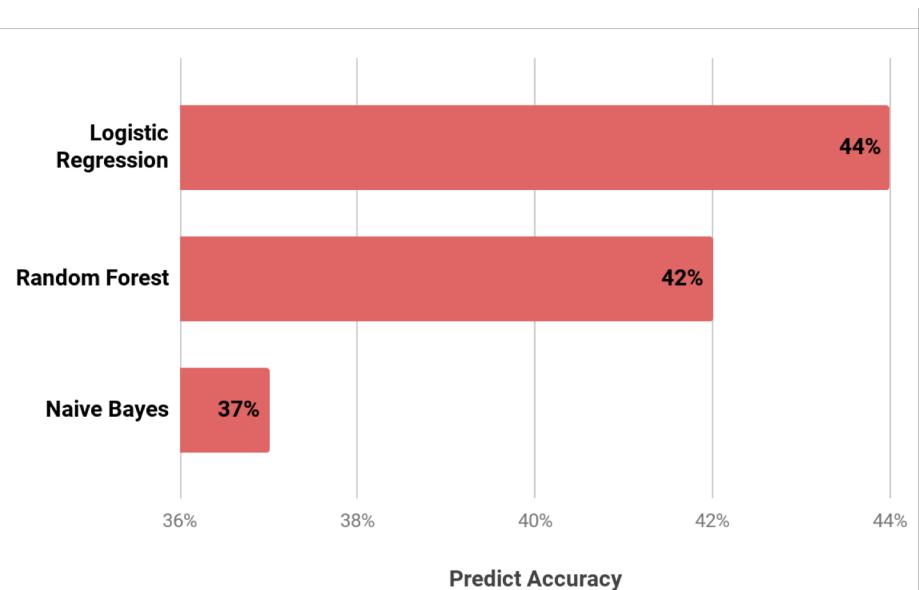
**5%**

Buffet, Pour,  
Cest, Mai, Nicht,  
Station, Plu, Trè,  
Dan, Sehr

**1%**

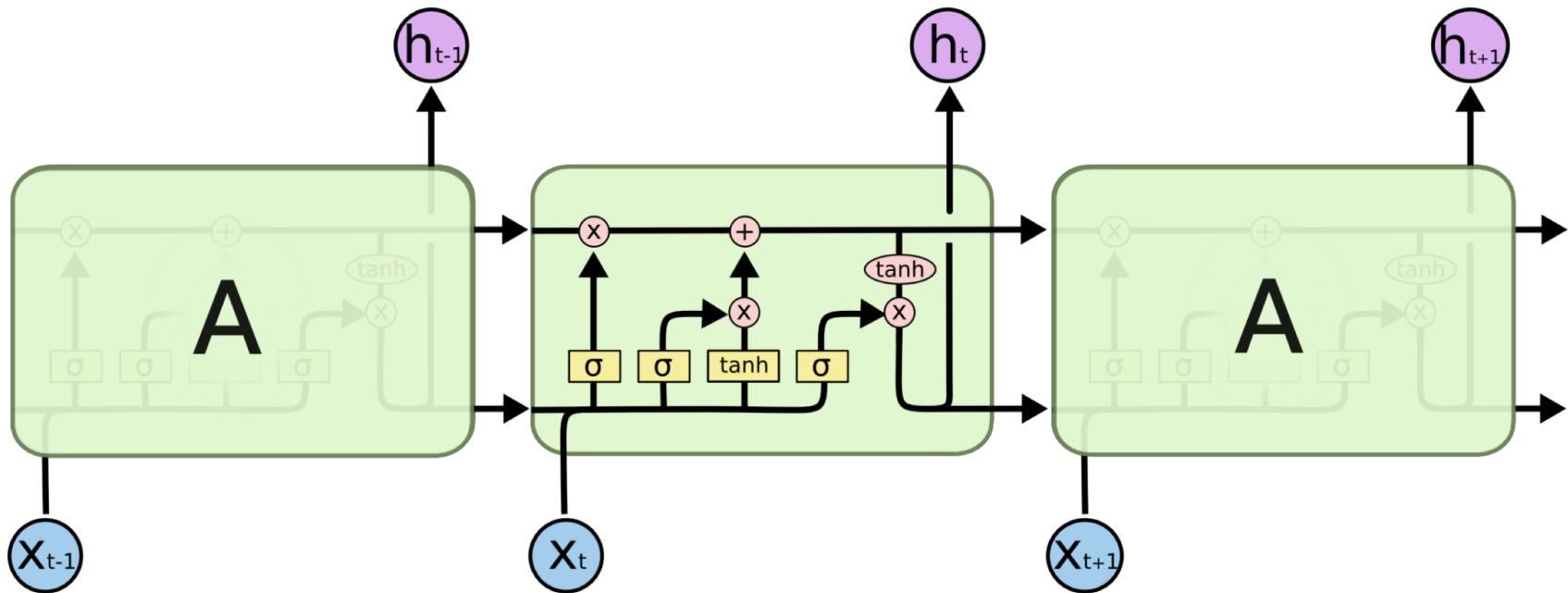
Order, Time,  
Food, Like, Wait,  
Would, Table,  
Even, Came,  
Didn't

# LDA Topics used for Machine Learning



---

# LSTM



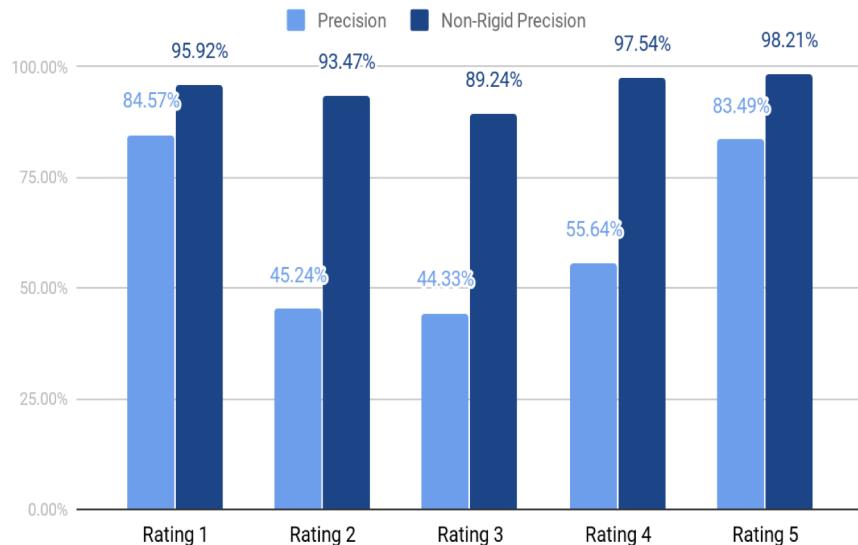
# LSTM

01	Vocabulary Size	<ul style="list-style-type: none"><li>• <b>50K most frequently used words</b></li></ul>	01	Vocabulary Size	<ul style="list-style-type: none"><li>• <b>80K most frequently used words</b></li></ul>
02	Training Dataset	<ul style="list-style-type: none"><li>• <b>25K reviews for training</b></li></ul>	02	Training Dataset	<ul style="list-style-type: none"><li>• <b>1 million reviews for training</b></li></ul>
03	Review Trim	<ul style="list-style-type: none"><li>• Maximum of <b>200 words</b> for each review</li></ul>	03	Review Trim	<ul style="list-style-type: none"><li>• Maximum of <b>300 words</b> for each review</li></ul>

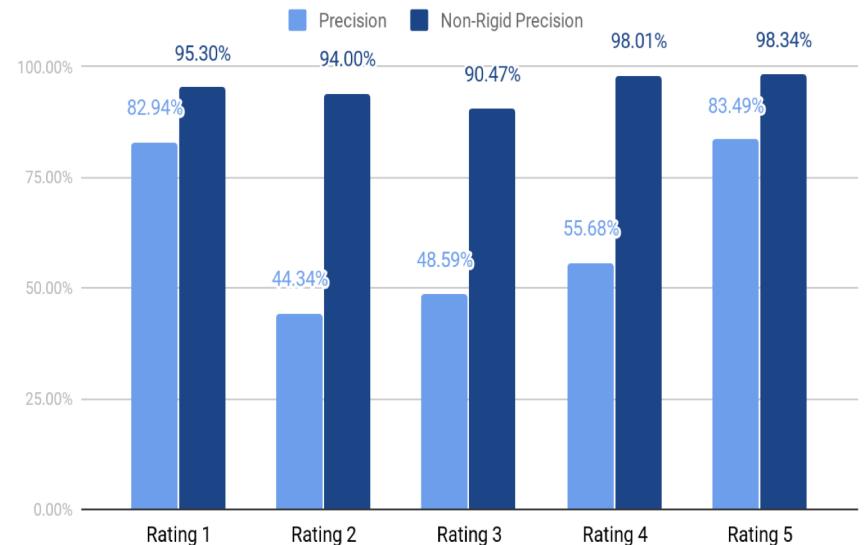


# Performance Evaluation

## Simple Network



## Complex Network



---



# References

- [1]. Alexandr, B. (n.d.). Rating prediction with sentiment analysis.
- [2]. Asghar, N. (n.d.). Yelp Dataset Challenge: Review Rating Prediction.
- [3]. Ganu, G., Elhadad, N., & Marian, A. (n.d.). Beyond the Stars: Improving Rating Predictions using Review Text Content.
- [4]. Jong, J. (n.d.). Predicting Rating with Sentiment Analysis.
- [5]. Kavousi, M. (n.d.). Estimating the Rating of Reviewers Based on the Text.

# Q & A

