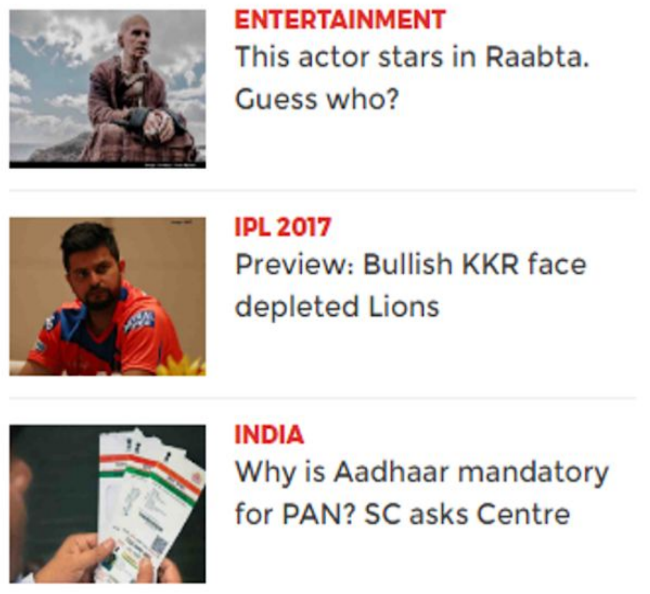
Improving Naive Bayes Prediction Models

Presented by Guangyan Cai, Zhaoyi Huang, Xiaoxin Xu
Under the supervision of Jerzy Lewak, Professor Emeritus, *University of California, San Diego*

Motivation

Email Spam Detection



ENTERTAINMENT
This actor stars in Raabta. Guess who?

IPL 2017
Preview: Bullish KKR face depleted Lions

INDIA
Why is Aadhaar mandatory for PAN? SC asks Centre

News Categorization

Introduction

This project aims to build a Text Classification model, one of the major topics in Machine Learning. Text Classification refers to identifying the category, or class, a piece of text belongs to using its content. Some applications involve filtering out spam emails and detecting abusive content, methods widely used by Facebook and Twitter nowadays.

Naive Bayes Classifier is a simple yet effective classifier for Text Classification problems. Although many new algorithms, like Neural Network, have been developed and employed, Naive Bayes Classifier remains one of the most popular algorithms due to its simplicity and relatively good accuracy.

The goal is to improve Naive Bayes Classifier by using **phrases** instead of **words**. We will experiment and test our classifier using data by National Highway Traffic Safety Administration (NHTSA). The data contains over one millions complaints about malfunctioning vehicles. Our objective is to predict whether a vehicle will be recalled in the future by classifying complaints

Vehicle Recall Data Representation

- Complaint

Complaint ID	Complaint Description	Vehicle Kind
268713	SEAT JAMS DUE TO FLOOR FLEXING. THE PROBLEM WA...	ford - bronco - 1989
268718	SEAT JAMS DUE TO FLOOR FLEXING. THE PROBLEM WA...	ford - bronco - 1989
717436	RIGHT REAR LIGHTING SYSTEM TO INCLUDE TURN SIG...	bmw - 3 series - 2003
717437	RIGHT REAR LIGHTING SYSTEM TO INCLUDE TURN SIG...	bmw - 3 series - 2003

- Recall

Recall Campaign Number	Vehicle Kind	Vehicle Kind Component
16V115000	open range - light - 2014	open range - light - 2014 - equipment
16V115000	open range - light - 2013	open range - light - 2013 - equipment
16V115000	open range - roamer - 2014	open range - roamer - 2014 - equipment
16V115000	open range - roamer - 2013	open range - roamer - 2013 - equipment

Methods

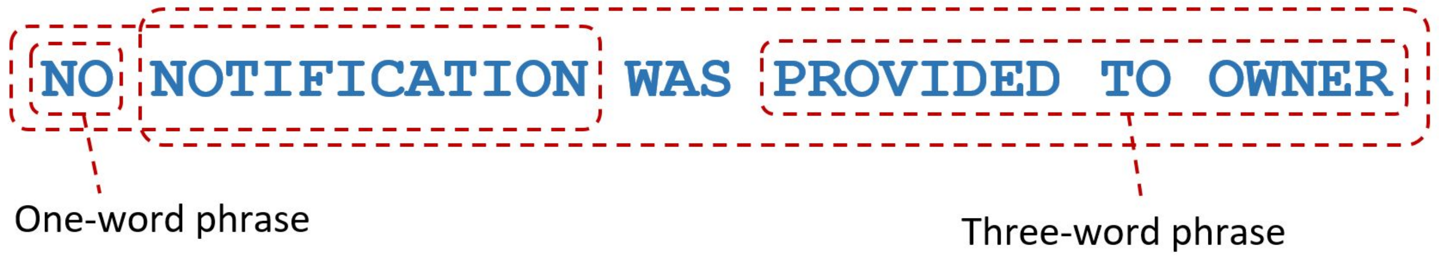
Naïve Bayes

- Naïve Bayes assumption:
  - Features (i.e. phrases) independently represent the likelihood of the class

$$P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y)$$
$$= P(X_1|Y)P(X_2|Y)$$

  - More generally:

$$\boxed{P(X_1 \ldots X_n|Y) = \prod_i P(X_i|Y)}$$



Calculate Phrase Score

- Extract all phrases from complaints up to 5 word length and omit those with relatively small frequencies(<10)
- Calculate phrase score: ln(#(recalls)/#(non-recalls))

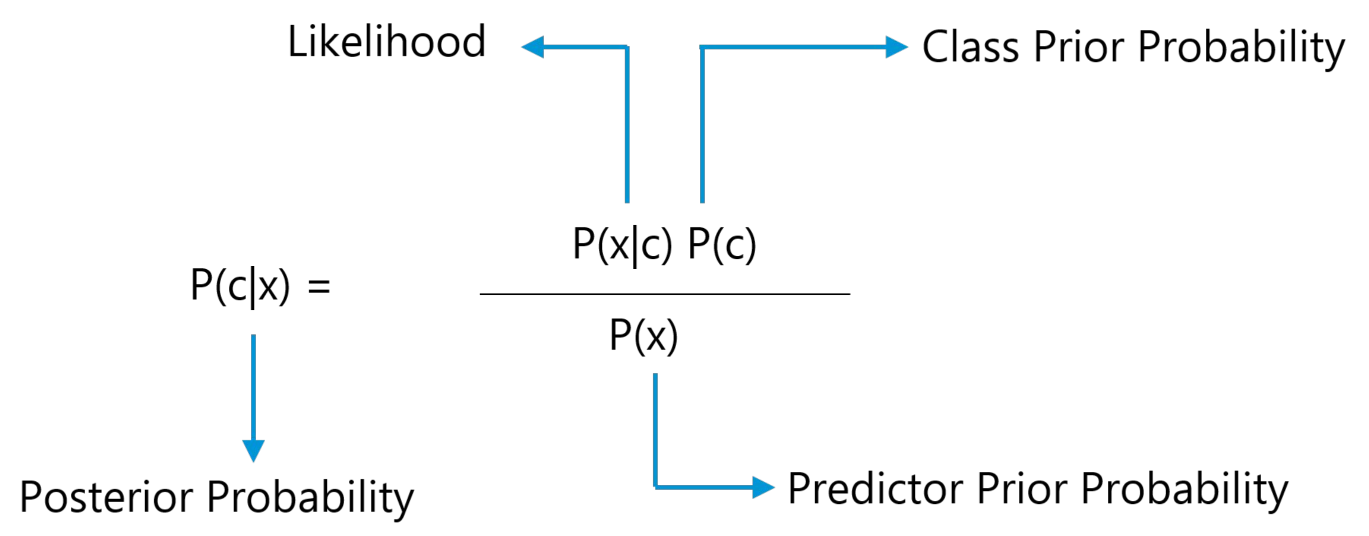
Calculate Complaint Score

- Calculate complaint score by adding phrase scores in that complaint

Build Prediction Model

- Predict recall/non-recall by determining the cut-off score of the complaints optimizing prediction accuracy using known results not used in training

Research Needed



- Explore the frequency threshold for phrases
- Investigate the optimal cut-off value of complaint scores of recall and non-recall
- Explore different ways of scoring complaints: use only phrases whose odds are greater than 1; use odds greater than 1 but only the highest odds phrase when it is a member of a family of multiple phrases; use all phrases.

Future Works

Other methods can be explored to calculate complaint scores based on phrase scores.

- Set a limit on the phrases by using the phrases with odds greater than a certain value
- Use the highest odds phrase of the family of phrases
- Test the consistency of prediction model by looking at a larger data set with 2016's data

References

- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Text Classification and Naive Bayes, Introduction to Information Retrieval*, Cambridge University Press. 2008. 3. Jurafsky,
- Dan, and James H. Martin. "Chapter 4: N-Grams" *Speech and language processing*. Pearson Education, 2014.