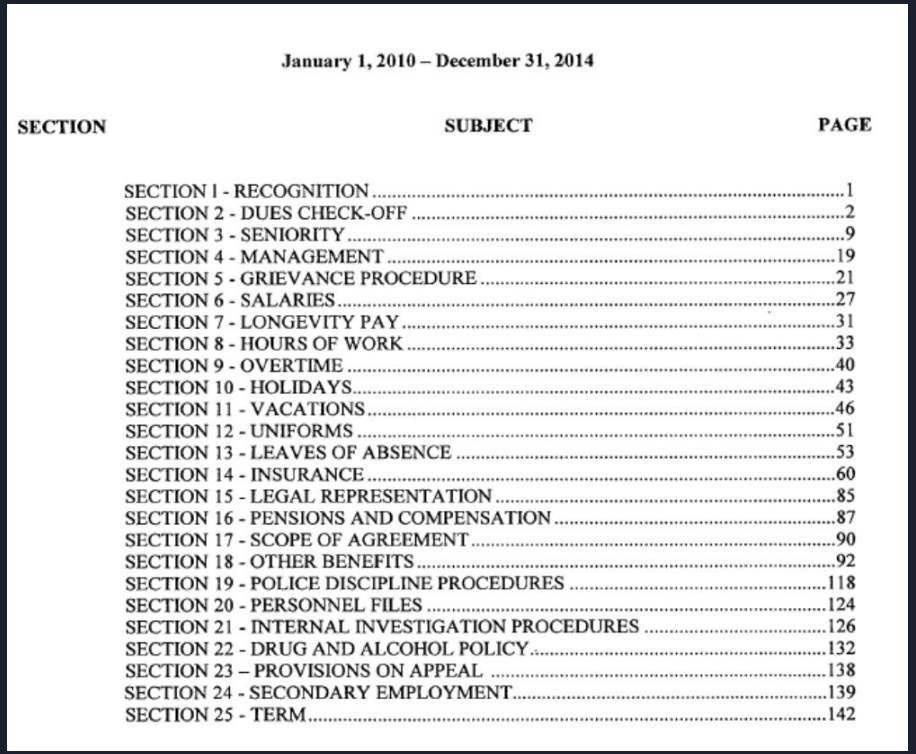
Police Union Contract Misconduct Complaint Detection system

Team - AngryNerds

Wang, Zian (email: ziw42@pitt.edu)
Gupta, Sonal (email: sog26@pitt.edu)
Zheng, Shuo (email: shz113@pitt.edu)

Introduction



January 1, 2010 – December 31, 2014

SECTION	SUBJECT	PAGE
SECTION 1 - RECOGNITION		1
SECTION 2 - DUES CHECK-OFF		2
SECTION 3 - SENIORITY		9
SECTION 4 - MANAGEMENT		19
SECTION 5 - GRIEVANCE PROCEDURE		21
SECTION 6 - SALARIES		27
SECTION 7 - LONGEVITY PAY		31
SECTION 8 - HOURS OF WORK		33
SECTION 9 - OVERTIME		40
SECTION 10 - HOLIDAYS		43
SECTION 11 - VACATIONS		46
SECTION 12 - UNIFORMS		51
SECTION 13 - LEAVES OF ABSENCE		53
SECTION 14 - INSURANCE		60
SECTION 15 - LEGAL REPRESENTATION		85
SECTION 16 - PENSIONS AND COMPENSATION		87
SECTION 17 - SCOPE OF AGREEMENT		90
SECTION 18 - OTHER BENEFITS		92
SECTION 19 - POLICE DISCIPLINE PROCEDURES		118
SECTION 20 - PERSONNEL FILES		124
SECTION 21 - INTERNAL INVESTIGATION PROCEDURES		126
SECTION 22 - DRUG AND ALCOHOL POLICY		132
SECTION 23 – PROVISIONS ON APPEAL		138
SECTION 24 - SECONDARY EMPLOYMENT		139
SECTION 25 - TERM		142

- Contracts are usually dispersed over different websites - tons of contracts and sessions that are hard to navigate.
- Many citizens are left in dark about how to approach police reforms

Motivation

- Socially Meaningful
- Interesting
- Powerful
- decrypted

What we are trying to achieve ?

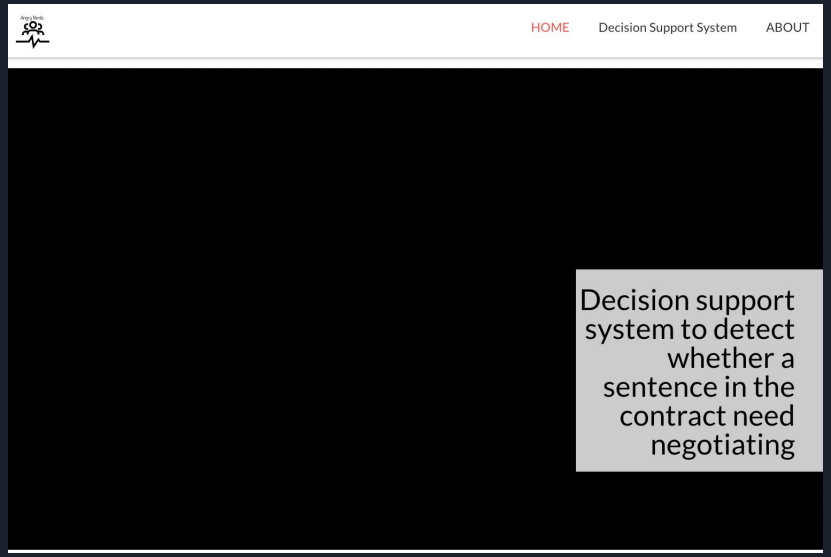


Fig 1. Website For Decision Support System

- Analyze the contracts from police departments.
- Discover problematic sentences and clauses.
- Categorize the problematic sentences.
- Build a website gives users a way to search for a sentence or select the sentence they feel confused about in a user-friendly way and displays the category of the sentence.

Libraries and Data we used

Main Libraries : stringdist, RTextTools (A supervised learning package for text classification)[1], R Shiny[2]

Data: Contracts data from “Police Union Contract Project” project[3].

- 87 police contracts from the 100 largest U.S. cities.
- Human-annotated data[4] as the ground truth to train the model.

Sentence	Category
Human Resources Department files are a …	Erases misconduct records
If the questioning is mechanically recorded …	Gives officers unfair access to information
…	…

Data Exploration

- Number of words increase :

  Longest Contract had 83,670 terms.

  Shortest Contract : 427 words

  Mostly in range - 15,000- 40,000 words

- Number of stop words also increase nearly.

  Some contracts had more than 20,000 stopwords.

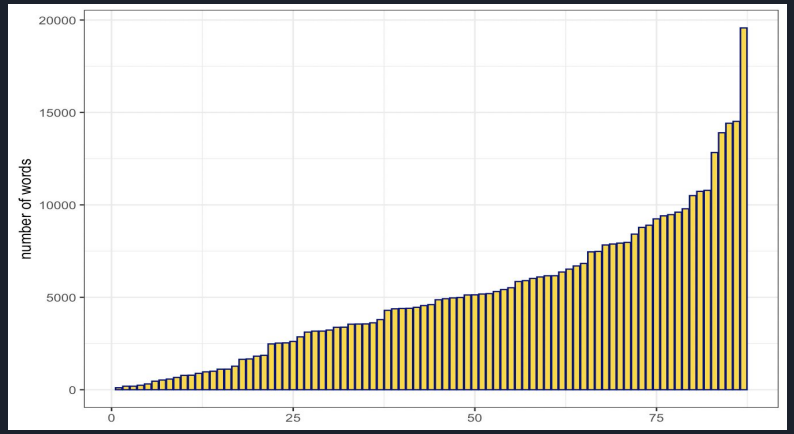


Fig 2. Number of Words

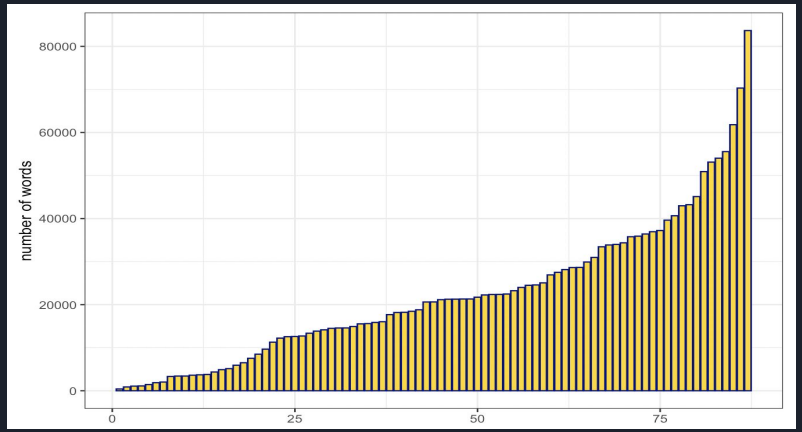


Fig 3. Number of Stop Words

Data Preprocessing(1)

Step-1: Convert all the pdf file into txt file

- Using Adobe Acrobat to convert into docx
- Using online converter tool to convert docx into .txt files

Step-2 : Read all the contracts into the R memory.

- Read a txt file one by one.
- Store sentences of length range(60, 500) in ‘text’ vector.

Data Preprocessing(2)

Step-3: Load the human-annotated data file (Ground Truth File)

- Get the `language` and `category` columns to define labels for sentences in `text` vector.

Step-4: Define the labels for problematic sentences

- Define a `label` vector to hold labels/ categories for sentences in `text` vector.
- `stringsim` function of `stringdist` library is used to calculate similarity score between sentences in `text` vector and `language` vector.
- If similarity score is greater than 0.85, the corresponding category is assigned to sentence otherwise it is a problematic sentence.

Data Preprocessing(3)

Step-5: Text Preprocessing in one Step -

Using `RTextTools` library: create_matrix() function to create term-document matrix

- Remove stop words
- Remove Numbers
- Stem the words
- Remove the sparse terms

```
```{r}
doc_matrix <- create_matrix(text, language="english", removeNumbers=TRUE,
                            stemWords=TRUE, removeSparseTerms=0.998)
```
```

Challenges in Data Preprocessing

- Labelling the Ground Truth (Human Annotated Data) with the text corpus

  -> Used `stringsim` function of `stringdist` library to calculate similarity score and decide

Data Modeling

Supervised

- We have the ground truth

Classification

- We need to classify the sentences into different problem categories

So what we got?

SVM (Support Vector Machines)

SLDA (Supervised Latent Dirichlet Allocation)

Boosting, Bagging, Random Forest, Neural Network

Decision Tree

Workflow

Tidy the data

Build the models

Evaluate the performances

Tidy the data

Problem 1:

Highly imbalance

1: 63
2: 55
3: 143
4: 303
5: 20952
6: 95
7: 294

Random over-sampling[4]

Tidy the data

Problem 2:
Test data sampling

Only over-sample
the training data

Build the models

Problem:

Too large dataset

Removing sparse terms:

Allow 0.998 sparsity:

Performance Evaluation(1)

Model	Precision	Recall	R-score
SVM	0.5400000	0.6585714	0.5857143
SLDA	0.2171429	0.6185714	0.2628571
Boosting	0.1571429	0.3557143	0.1400000
Bagging	0.5414286	0.6257143	0.5742857
Forest	0.6357143	0.6600000	0.6371429
Tree	0.1442857	0.1714286	0.1457143
Neural Network	0.2800000	0.5057143	0.3185714

Table 1: Performances of all the models

Performance Evaluation(2)

	Coverage	Recall
n>=1	1.00	0.81
n>=2	1.00	0.82
n>=3	1.00	0.82
n>=4	0.99	0.83
n>=5	0.96	0.87
n>=6	0.85	0.98
n>=7	0.51	0.99

Table 2: Ensemble agreement coverage and recall[5]

Performance Evaluation(3)

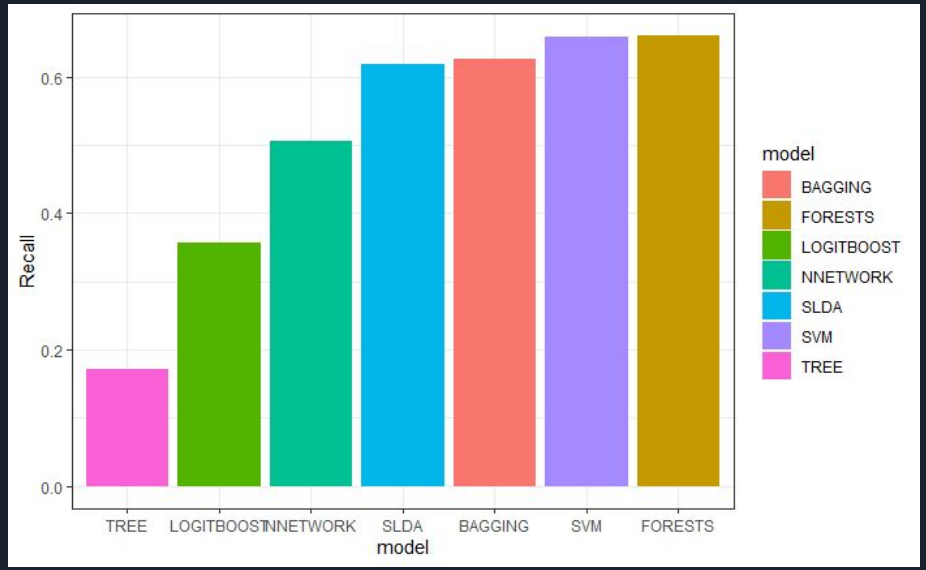


Fig 4: Recall of the models

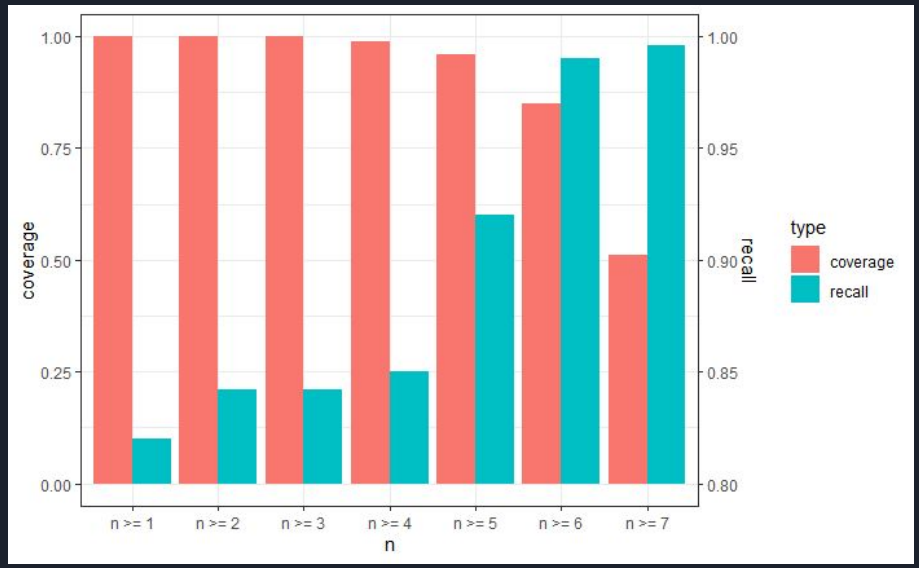


Fig 5: Coverage and recall at several ensemble cut-points

Decision Support System(1)

- We use text mining and model training to categorize our document sentences
- put data in the website, provide users two ways (select or search box) to use the system to predict problematic sentences.
- CLICK get started to get start!

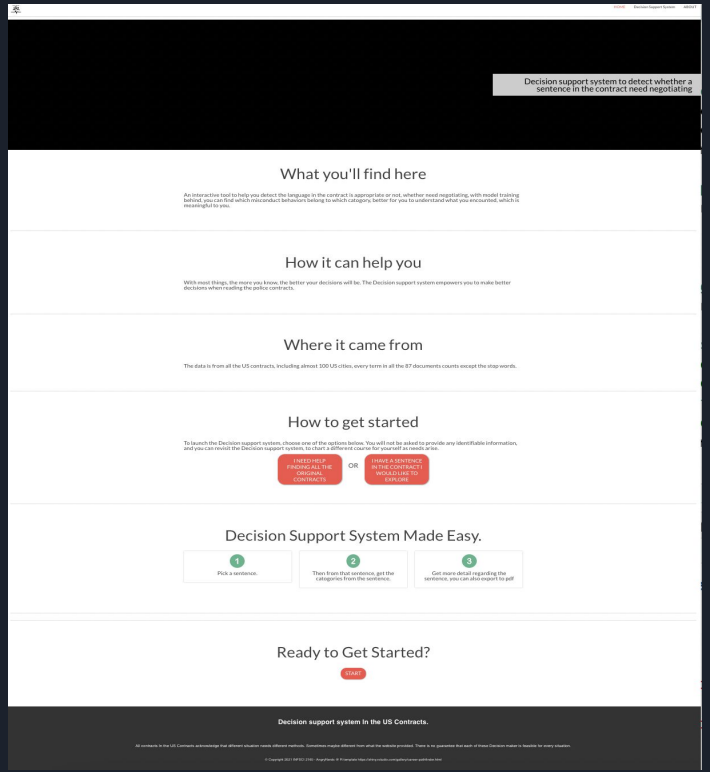


Fig 6: Home page of Decision Support system

Decision Support System(2) - Steps

- Step 1 : Input a sentence or select the sentence which you want
- Step 2: Given the sentence, system predicts the category of the sentence
- Step 3: Ask for help based on the category

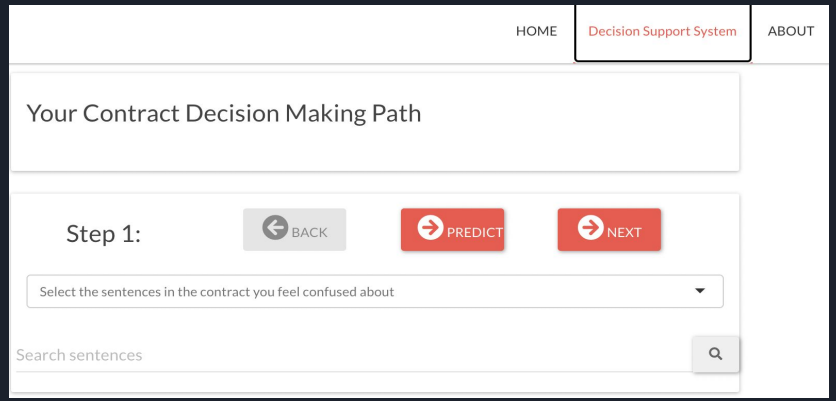


Fig 7: Decision Support System Page - Step 1

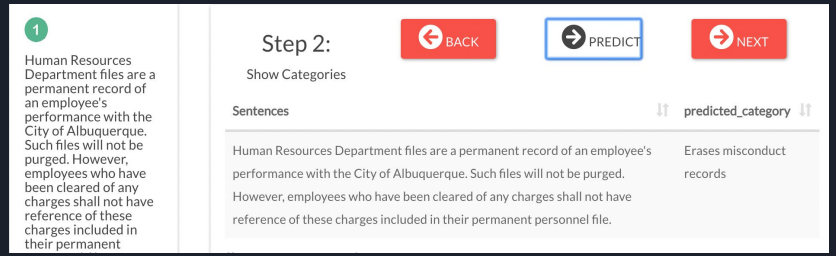


Fig 8: Decision Support System Page - Step 2

https://shz113.shinyapps.io/Decision-Support-System/

Our Creative Approach

- ★ Deal with the real world problems - USE Shiny app to create UI system, providing a user-friendly way to solve the problem
- ★ Text mining based on all the pdf contracts, create a way to deal with the pdf contracts
- ★ Model training can get high recall results, can predict any sentence regarding the police contracts and output good results

Fig 9: About page