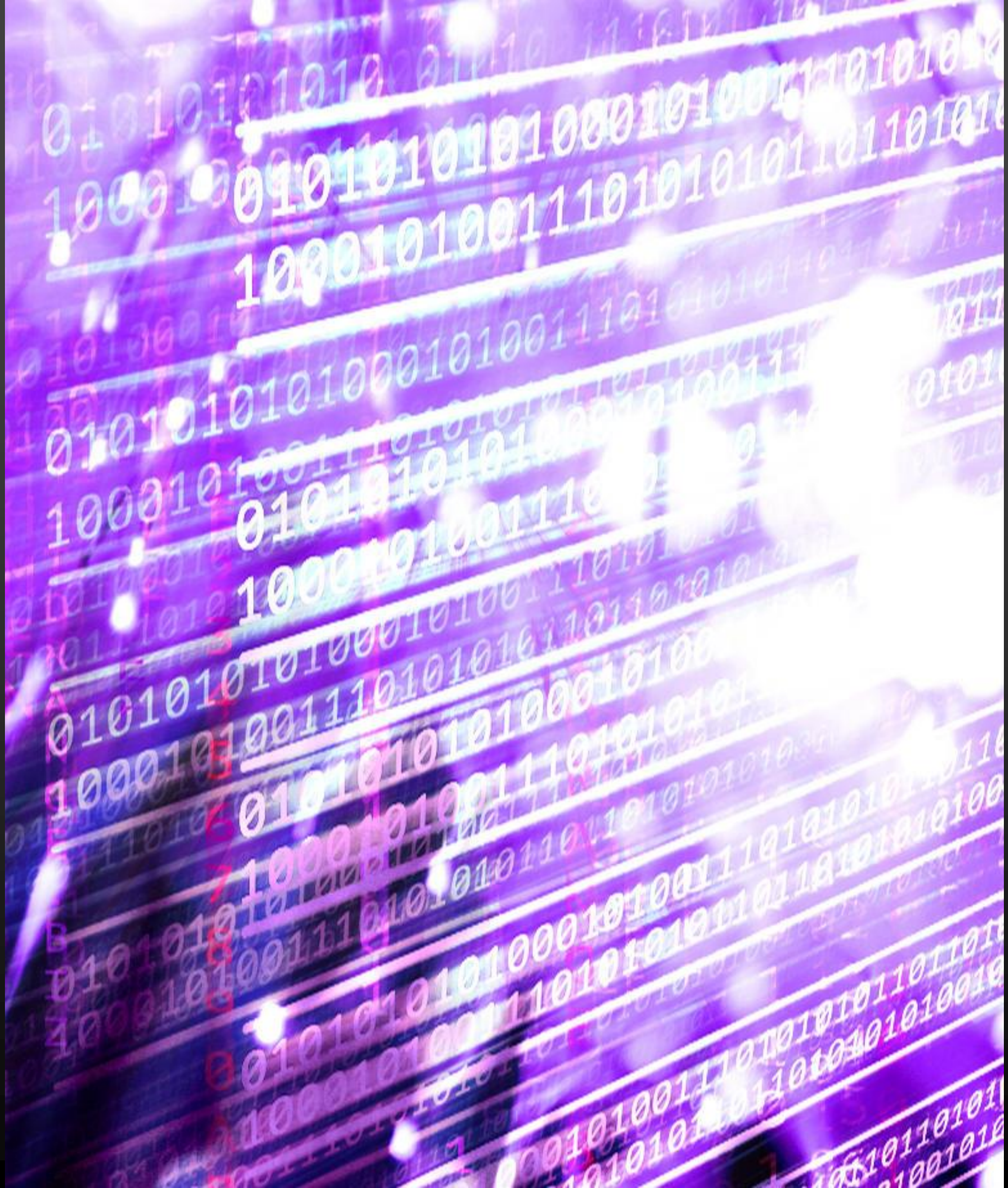
Text Classification on Public Unstructured Healthcare Data

Francisco Salas



Sections

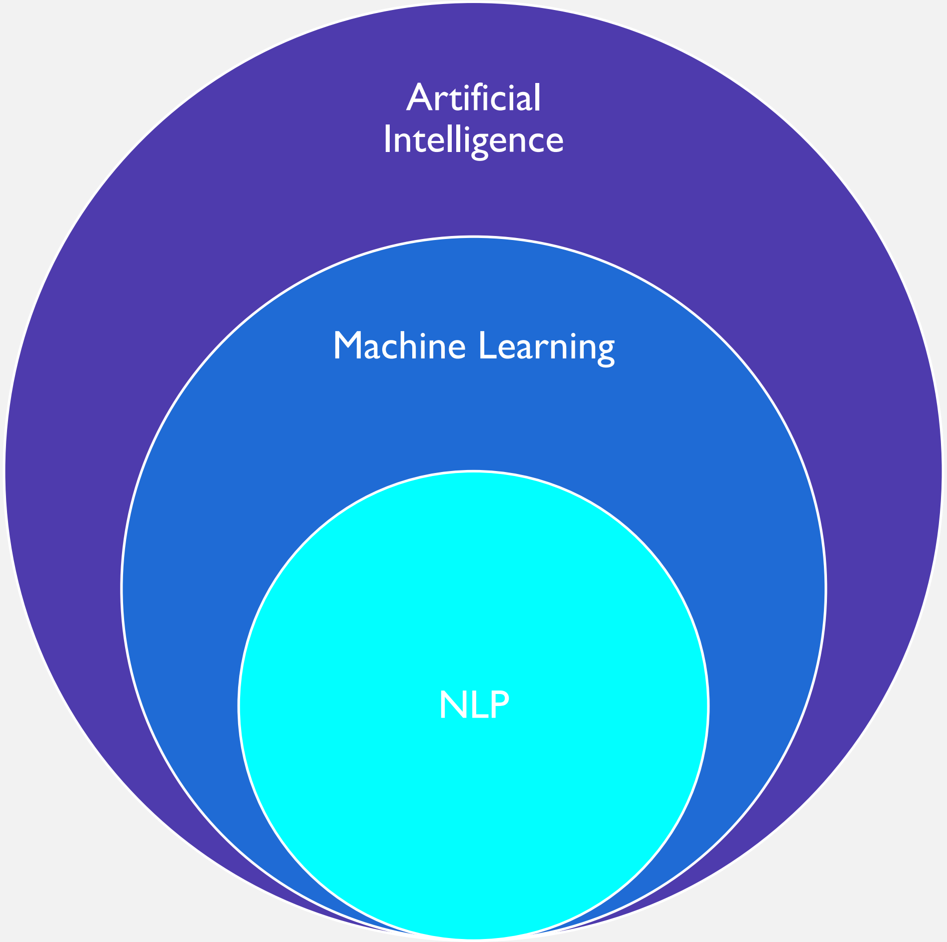
Subtitle lorem ipsum dolor

- NLP & Healthcare Overview
- Problem
- Data Acquisition & Cleaning
- Models
- Results

Definitions

- Artificial Intelligence : (AI) Branch of Computer Science that deals with the study of “intelligent agents”; any device that perceives its environment and takes action to maximize its change of achieving its goal.
- Machine learning (ML) : Subfield of AI that explores the study and construction of algorithms that can learn from and make predictions of data.
- Natural Language processing (NLP): Area of AI that deals with the interactions between computers and human ***natural*** languages

(over)Simplified Diagram



Artificial Intelligence

Machine Learning

NLP

Key applications areas of NLP

- Sentiment analysis : Identifying subject information in the text whether it conveys judgment, opinion or reviews etc.
  - Polarity detection
- Text classification : Assign predefined tags to sections of text.
  - Binary classifier
  - Multilabel classification
- Topic Modeling: Extract hidden topics form large volume of text
  - Latent Dirichlet Allocation (LDA)

Everyday NLP Applications

- Spam filter : Gmail
- Autocomplete : SwiftKey
- Enhance grammar check : Grammarly
- Language detection : Google Translate

Healthcare Data

Structured

- Patient demographics
- Medication list
- Family health history
- Lab results

Unstructured

- Physician notes on patient results
- Imaging test results
- PDF paper records
- **Patient survey/opinions**

HIPPA

- Health Insurance Portability and Accountability Act
  - Portability of Insurance
  - Protection and Privacy of Healthcare Information
  - Standardization and Efficiency in Healthcare Data
  - Prevention of Discrimination and Fraud

Public Healthcare Data: Unstructured

Specialized

- WebMD
- RateMds
- Vitals
- ZocDoc

Third party

- Angie's List
- Facebook
- Yelp
- Twitter

Problem

Problem

- The need for analyzing unstructured healthcare data is increasing
- Tools used for public data can be implemented for private data
- Text Classification
  - Supervised categorical
  - Predicting labels from text data

Francisco Salas

Data Acquisition & Cleaning

Pubic Dataset

https://www.yelp.com/dataset

- Yelp Open Dataset
- 5,996,996 reviews
- 188,593 business
  - **Including Healthcare locations !!**
- > 1.4 million business attributes

Yelp Dataset Files

- `yelp_academic_dataset_business.json`
  - Contains business data including location data, attributes, and categories
- `yelp_academic_dataset_review.json`
  - Contains full review text data including the user_id that wrote the review and the business_id the review is written for.
- `yelp_academic_dataset_user.json`
  - User date including the user's friends mapping and all the metadata associated with the user
- `yelp_academic_dataset_checkin.json`
  - Chekings on a business
- `yelp_academic_dataset_tips.json`
  - Tips written by a user on a business
- `yelp_academic_dataset_photo.json`
  - Contains photo data including the caption and calssification

Yelp Dataset Description

yelp_academic_dataset_business.json

Name	Description
business_id	Unique id
categories	Array of string of business categories
name	Name of a business
state	Two character code of state
…	..

yelp_academic_dataset_review.json

Name	Description
review_id	Unique id review
user_id	Unique id user
business_id	Unique business id
stars	Star rating
date	Date of review
text	String review itself
…	…

Data Cleaning steps

Select datasets	• `yelp_academic_dataset_business.json` • `yelp_academic_dataset_review.json`	
Drop null values	• Removes rows with null values • *`df.isnull()`*	
Filter *`df.categories`* unique to healthcare	• family practice • urgent care • obstetricians & gynecologists • cosmetic surgeons • internal medicine	• dermatologists • surgeons • ear nose & throat • psychiatrists • …
Filter *`df.state`* unique to US states	• AZ • NV • PA • NC • OH	• IL • WI • CA • OR
Merge with datasets with similar key *`pd.merge(business,review, on='business_id')`*	• `business_id` • `categories` • `name` • `state` • `cool` • `date`	• `funny` • `review_id` • `stars` • `text` • `useful` • `user_id`

Francisco Salas

Yelp Clean data results

Original data

Total reviews	5.9 million
Total unique business	188 K

Clean data

Total reviews	44,918
Total unique business	3062

Models

Francisco Salas

Data Selection

Classification : Supervised Learning

column	Description
text	• String • 5000 character limit • User review about a specific business • *Features*
stars	• Integer • Value range [1,2,3,4,5] • User personal score for a given business • Labels
Health Business	• chiropractors • hospitals • family practices • obstetrician • diagnostic service • urgent care • physical therapy • mental health

Sentiment analysis :TextBlob

- Method to quantify qualitative data with some sentiment score, goal is to extract the emotion content of text
- Sentiment dictionary approach, Mapping words to sentiment values

Text Classification : *Supervised Learning*

Naïve Bayes

- Implements the naive Bayes algorithm for multinomially distributed data, and is one of the two classic naive Bayes variants used in text classification

Support Vector Machines

- Constructs a separating line called a hyperplane witch can be use for classification
- Margin maximizes the distance to the nearest point

Feature Extraction

Count Vectorizer

- Produces a bag-of-words representation from a document or corpus
- Convert a collection of text documents to a matrix of token counts
-

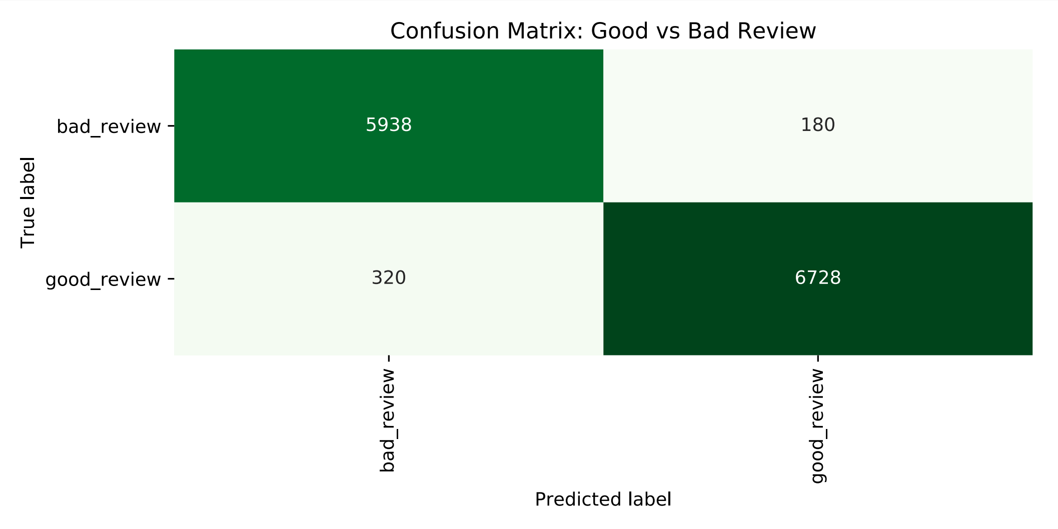
TF-IDF

- Term Frequency – Inverse Document frequency
- Useful for finding term that are important for the specific document and uncommon in the corpus as a whole.

Results

Naïve Bayes

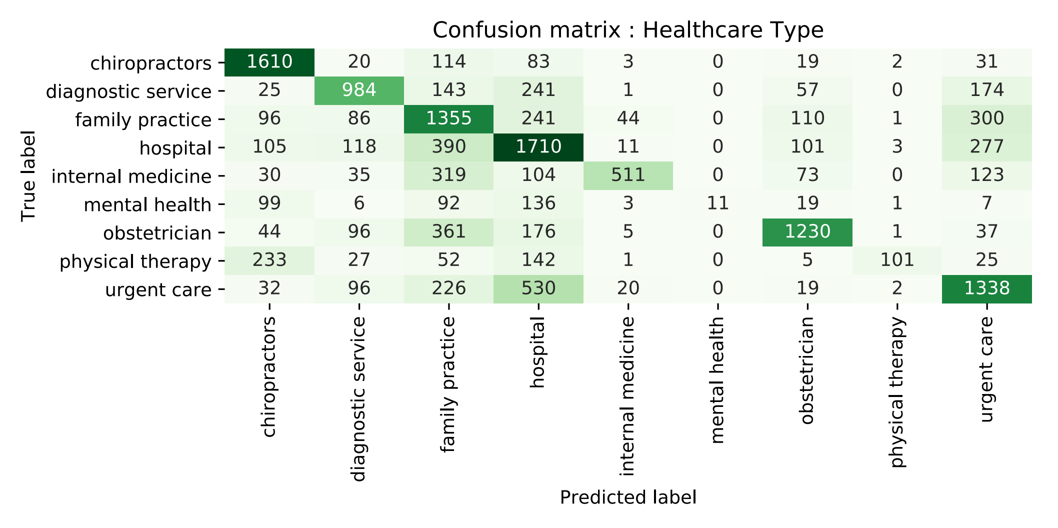
Text Classification : Star review



Confusion Matrix: Good vs Bad Review

True label \ Predicted label	bad_review	good_review
bad_review	5938	180
good_review	320	6728

Text Classification Healthcare Business



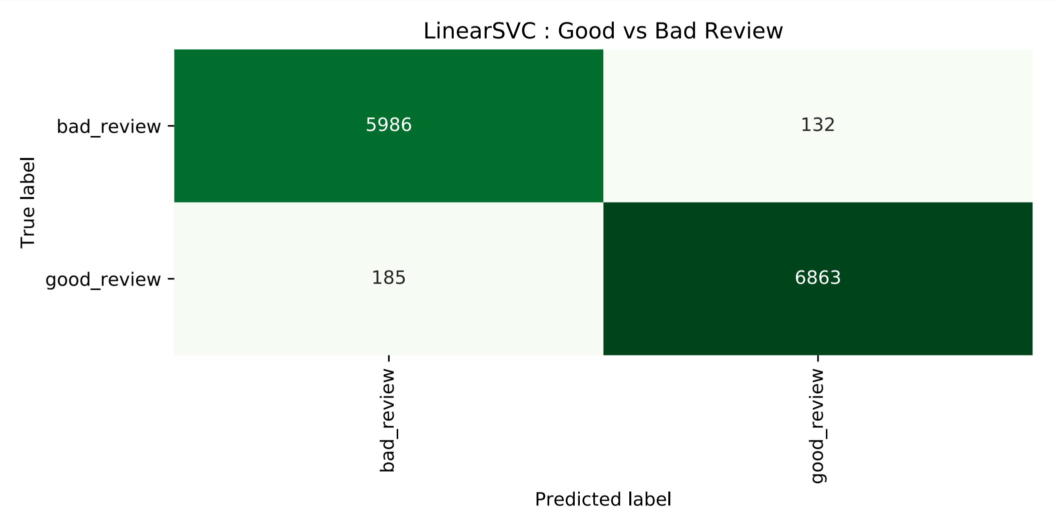
Confusion matrix : Healthcare Type

True label \ Predicted label	chiropractors	diagnostic service	family practice	hospital	internal medicine	mental health	obstetrician	physical therapy	urgent care
chiropractors	1610	20	114	83	3	0	19	2	31
diagnostic service	25	984	143	241	1	0	57	0	174
family practice	96	86	1355	241	44	0	110	1	300
hospital	105	118	390	1710	11	0	101	3	277
internal medicine	30	35	319	104	511	0	73	0	123
mental health	99	6	92	136	3	11	19	1	7
obstetrician	44	96	361	176	5	0	1230	1	37
physical therapy	233	27	52	142	1	0	5	101	25
urgent care	32	96	226	530	20	0	19	2	1338

Francisco Salas

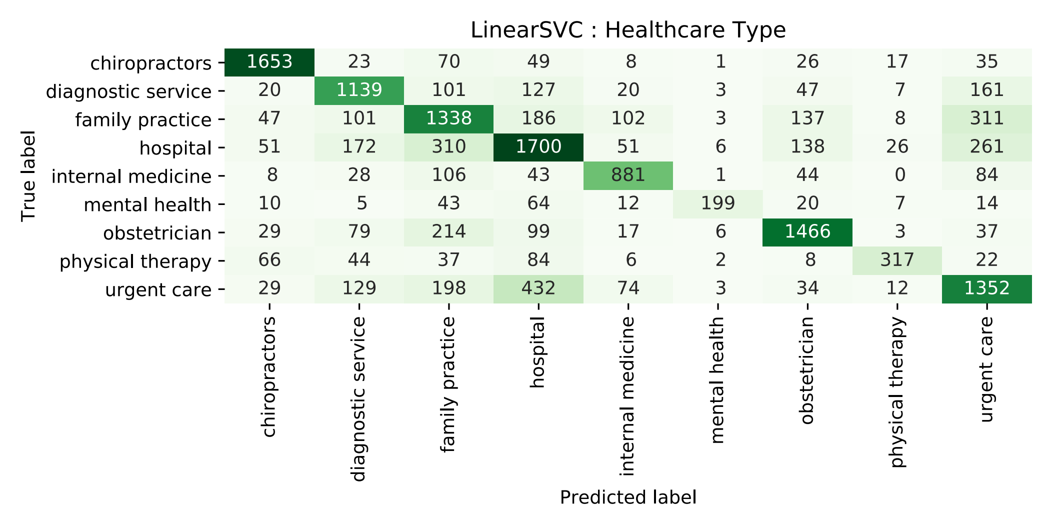
SVC

Text Classification: Star review


LinearSVC : Good vs Bad Review

True label \ Predicted label	bad_review	good_review
bad_review	5986	132
good_review	185	6863

Text Classification : Healthcare Business


LinearSVC : Healthcare Type

True label \ Predicted label	chiropractors	diagnostic service	family practice	hospital	internal medicine	mental health	obstetrician	physical therapy	urgent care
chiropractors	1653	23	70	49	8	1	26	17	35
diagnostic service	20	1139	101	127	20	3	47	7	161
family practice	47	101	1338	186	102	3	137	8	311
hospital	51	172	310	1700	51	6	138	26	261
internal medicine	8	28	106	43	881	1	44	0	84
mental health	10	5	43	64	12	199	20	7	14
obstetrician	29	79	214	99	17	6	1466	3	37
physical therapy	66	44	37	84	6	2	8	317	22
urgent care	29	129	198	432	74	3	34	12	1352

Thank You

Francisco Salas

✉ Frank.salas@gmail.com