IST-332 Final Project

NLP on Yelp's review data

Whoa!

Fayez Alharbi, Yuri Yu, Marium Mukhtar, Umesh Makhloga

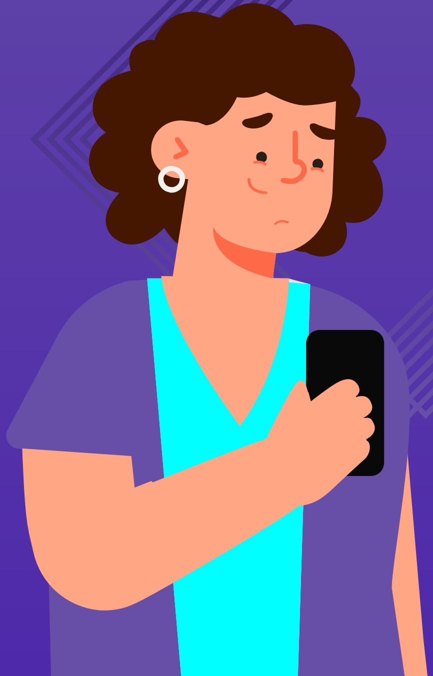


Table of **contents**

01 Introduction

Description of the problems
Summary of the overall project

Description of the problems

The main goal of this project is to collect review data from outside the Regional Health Care Plan (RHCP) resources and identify high-quality businesses.

Using customers' reviews to identify and measure the quality of the business instead of Yelp review ratings.

Summary of the overall project

We collected customers reviews about healthcare providers that provide services to Riverside and San Bernardino counties on Yelp.

We extract the feature sets from the review texts, Then train them to build our models.

Use the best model to help identifying good business

02

Corpus creation

Describe the steps for corpus creation
Summary statistics of the corpus

Describe the steps for corpus creation

-Read csv to load all the reviews metadata
- Save the corpus
- Summary of statistics of the corpus

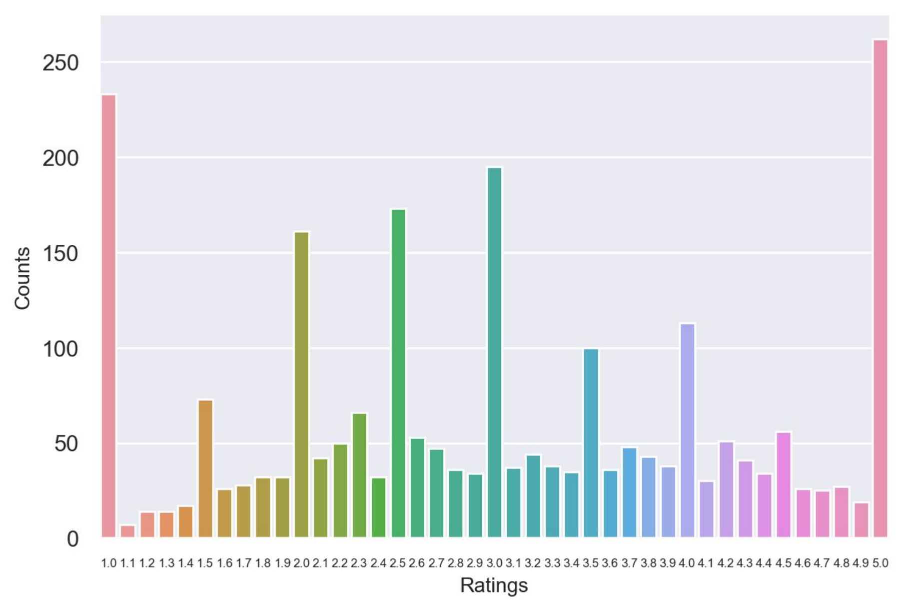
Business ID	rounded_rating	doctorID	Name	Business Category	review_content
chuang-t-hung-md-upland	2.7	101	Chuang T. Hung MD	Gastroenterologistgastroenterologist,	Best guy to check what's going on up there!!! ...
chuang-t-hung-md-upland	2.7	101	Chuang T. Hung MD	Gastroenterologistgastroenterologist,	This review does not reflect what I think of t...
chuang-t-hung-md-upland	2.7	101	Chuang T. Hung MD	Gastroenterologistgastroenterologist,	I have been having issues with my liver/stomac...

Summary statistics of the corpus

```
# A strongly skewed distribution. A small number of businesses have a relatively large number of reviews, while most businesses have relatively few.
# The max (633) is so much bigger than the mean (26) that it skews the statistics significantly. Maybe helpful is to note that the 50%
# percentile number of reviews is only 10 and the 25% percentile is 4, those counts are significantly less than the average.
# This might present some interesting challenges for machine learning as we will have significantly more data for some businesses then we do for others.
```

```
Business ID
24-7-care-at-home-westminster-2                                 7
4-ever-green-collective-riverside                               9
a-doctors-weight-loss-clinic-moreno-valley-2                   13
a-gobaud-orthopaedic-medical-clnc-and-bck-trtmnt-ctr-montclair  1
a-healing-within-palm-desert                                   14
                                                               ..
yusufaly-imdad-md-wildomar                                     19
yvonne-d-sylva-md-corona                                       49
zacher-judith-md-palm-desert                                    5
zeid-k-kayali-md-rialto-2                                      11
zosima-b-cariño-gateb-md-indio-2                                1
Name: Business ID , Length: 2468, dtype: int64
```

```
count    2468.000000
mean       25.827796
std        49.925814
min         1.000000
25%         4.000000
50%        10.000000
75%        26.000000
max       633.000000
Name: Business ID , dtype: float64
```



03

Text Preprocessing

Steps for text preprocessing all business reviews

Steps for Preprocessing

Tokenization and normalization
Contraction Expansion
Word_punct Tokenizer
nltk.pos_tag
Lemmatization
Checking for digits
Removing Punctuation
Removing words w/ less than two tokens
Checking for misspelling
Applying lexical diversity
Applying Frequency distribution
Getting a count for raw tokens, cleaned tokens and number of misspellings

	Username	Business ID	Business Name	Raw Review	Normalized Review	Raw Review Length	Raw Review Unique Token	Raw Review Lexical Diversity	Normalized Review Length	Normalized Review Unique Token	Normalized Review Lexical Diversity
0	Gregory P.	chuang-t-hung-md-upland	Chuang T. Hung MD	[Best, guy, to, check, what, is, going, on, up...	[best, guy, check, !!!, many, many, year, alwa...	31	27	1.148148	13	11	1.181818
1	Corvetta M.	chuang-t-hung-md-upland	Chuang T. Hung MD	[This, review, does, not, reflect, what, I, th...	[review, reflect, think, business, however, re...	135	75	1.800000	55	35	1.571429
2	Micky B.	chuang-t-hung-md-upland	Chuang T. Hung MD	[I, have, been, having, issues, with, my, live...	[issue, liver, stomach, couple, year, real, so...	296	162	1.827160	119	94	1.265957

	Business ID	Business Name	Business Category	Raw Review	Normalized Review	Raw Review Length	Raw Review Unique Token	Raw Review Lexical Diversity	Normalized Review Length	Normalized Review Unique Token	Normalized Review Lexical Diversity
0	24-7-care-at-home-westminster-2	24/7 Care At Home	Podiatristspodiatrists, Home Health Carehomehe...	[A, great, Home, health, service, located, rig...	[great, home, health, service, locate, right, ...	692	298	2.322148	289	182	1.587912
1	4-ever-green-collective-riverside	4 Ever Green Collective	Medical Centersmedcenters, Cannabis Clinicscan...	[4, EVER, GREEN, COLLECTIVE, 2781, Rubidoux, B...	[ever, green, collective, rubidoux, blvd, rive...	1374	579	2.373057	613	371	1.652291
2	a-doctors-weight-loss-clinic-moreno-valley-2	Doctor's Weight Loss Clinic	Doctorsphysicians,	[Doctor, Brysk, and, her, staff, is, great, ,...	[doctor, brysk, staff, great, recommend, frien...	957	397	2.410579	393	237	1.658228

04

Data Understanding

Steps taken for all business reviews

Identify any potential issues

Mixed Tokens

```
df['review_content_pre'][0]
```

```
['best',
 'guy',
 'check',
 '!!!',
 'many',
 'many',
 'year',
 'always',
 'recommend',
 'team',
 'thanks',
 'doc',
 '!!!']
```

Non-English Tokens

```
df['review_content_pre'][43430][-31:]
```

```
['evade',
 'responsibility非常不好的体验',
 '我老婆在这里生的孩子',
 '我的医生在36周就转了资料给医院',
 '但是医院从来没跟我们核实过资料信息',
 '我们提交了自己的保险',
 '可是因为医院的不负责任',
 '没有告诉我们他们是否接受我的保险',
 '没有履行他们的责任',
 '连最基本的资料都不核对',
 '这是对产妇的不负责',
 '也是他们工作的不重视',
 '这种连基本工作都做不好的医院',
 '你们敢把生命交给他们吗',
 '不出事没问题',
 '出事了那有多少麻烦等着你',
 '因为他们不告知我们的保险不接受',
 '就同样接受生子预定的情况下',
 '给我们造成了他们是接受我的保险的误会',
 '现在保险公司付款网外了的部分了',
 '医院还要找我们收取1万3美金',
 '想和医院协商',
 '医院的工作人员及其的不耐烦',
 '对于他们的工作失误',
 '也不管不问',
 '这种黑心医院',
 '建议不要来这里',
 '经济几次协商',
 '他们不退让',
 '不承认自己有失职',
 '责任推卸的一干二净']
```

Misspells

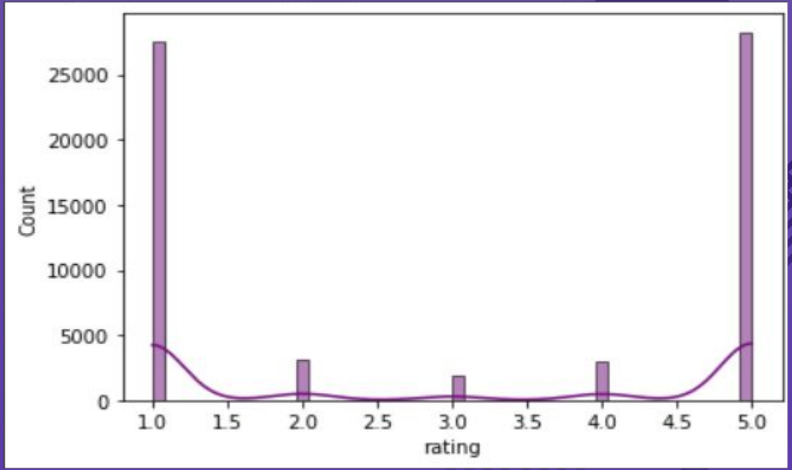
```
np.mean(df_test['misspells'])
```

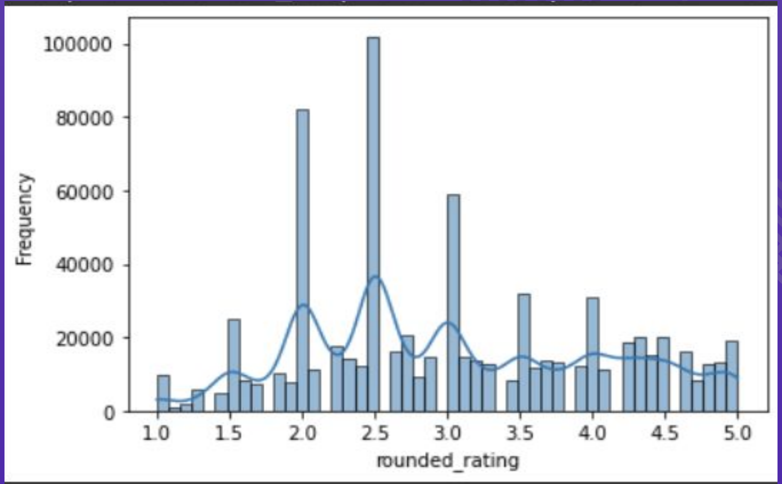
```
1.0516135105031141
```

- df.describe()
- df.drop() (null values, unneeded columns)
- Removing special characters, removing non english reviews [back to preprocessing]
- Frequency, Count of [‘rounded_rating’, ‘rating’,]
- [‘business lexical diversity’]
- df.groupby([‘Business Category’, ‘date_of_review), mean, max

```
Business Category                                                     date_of_review
Addiction Medicineaddictionmedicine, Counseling & Mental Healthc_and_mh,  1/14/2013        5.000
                                                                      1/14/2018        5.000
                                                                      1/15/2014        1.000
                                                                      1/15/2019        1.000
                                                                      1/17/2020        1.000
                                                                                        ...
Walk-in Clinicswalkinclinics, Urgent Careurgent_care,                 9/5/2017         3.000
                                                                      9/5/2018         2.000
                                                                      9/8/2017         1.000
                                                                      9/8/2020         5.000
                                                                      9/9/2013         1.000
Name: rating, Length: 44697, dtype: float64
```

Visualization





05

Sentiment Analysis

Steps of Sentiment Analysis

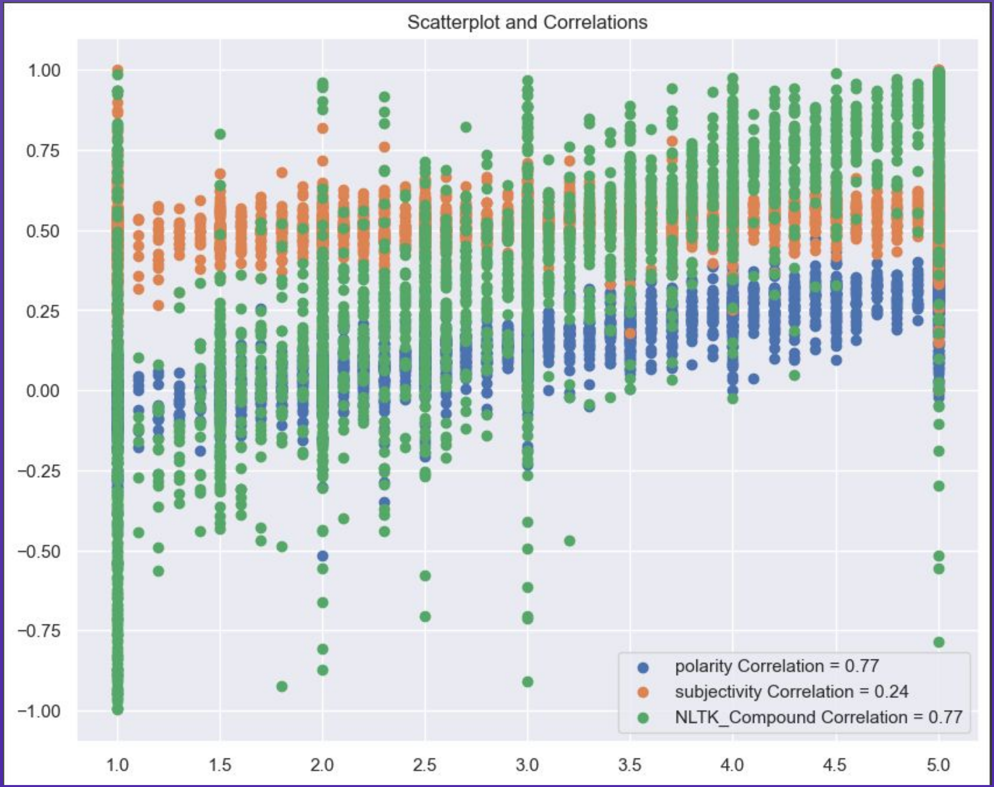
Concise explanation

Sentiment Analysis steps

- Aggregation
- Finding the scores of
  - polarity
  - subjectivity
  - NLTK_Compound

polarity	subjectivity	NLTK_Compound
0.405979	0.642869	0.962729
0.391728	0.585269	0.799167
0.330723	0.580852	0.664585
-0.300000	0.600000	-0.440400
0.225908	0.596353	0.538986

Correlations in sentiment analysis



06

Topic Modeling

Steps text preprocessing for all business reviews

Topic Modeling steps

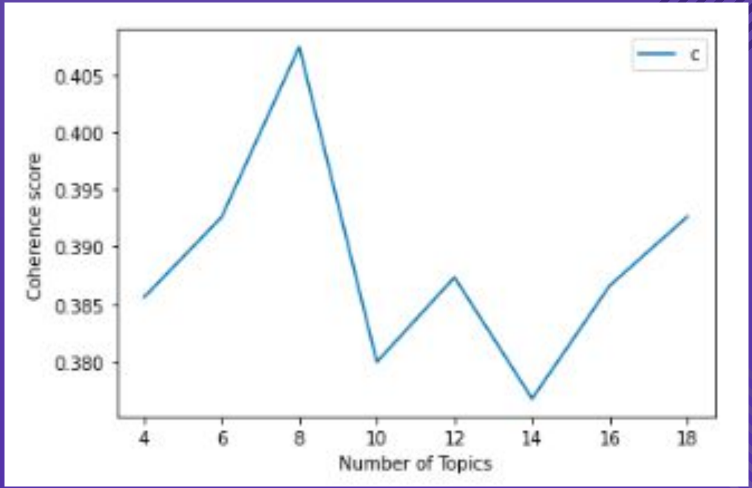
- Dictionary and corpus for Gensim
- Create LDA topic models
- Selecting topic numbers based on coherence scores
- Adding a label to each topic in the best topic mode
- Adding a visualization on the best topic model

Best topic model

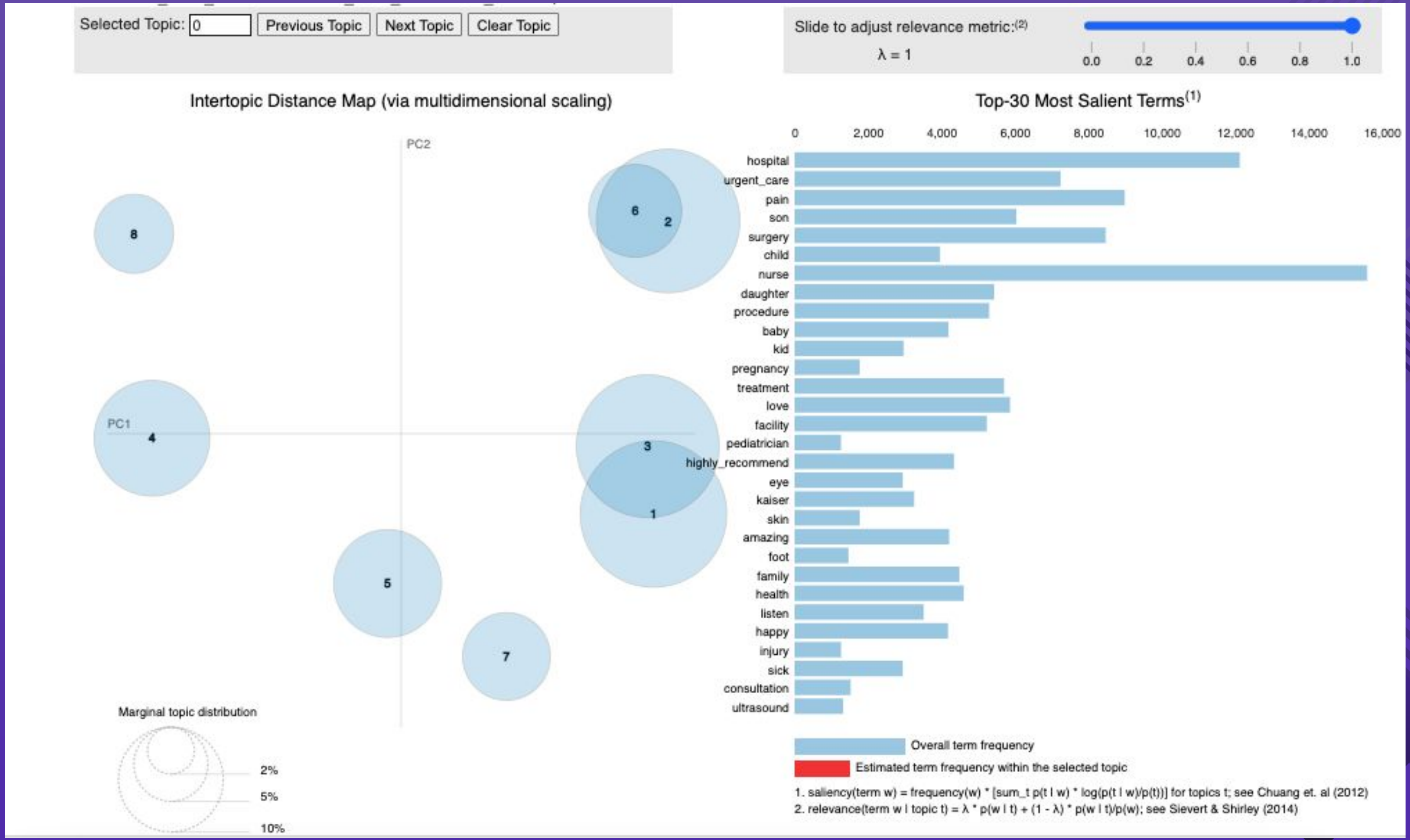
Top words in each topic for exp-1

```
for index, value in lda_model1.print_topics(num_topics=8, nu
    print('Topic ', index, ':', value)
```

```
Topic  0 : 0.028*"son" + 0.025*"child" + 0.022*"daughter" +
Topic  1 : 0.007*"pay" + 0.006*"send" + 0.005*"receptionist"
Topic  2 : 0.024*"urgent_care" + 0.008*"nurse" + 0.007*"test
Topic  3 : 0.010*"family" + 0.009*"health" + 0.009*"listen"
Topic  4 : 0.022*"surgery" + 0.017*"procedure" + 0.011*"eye"
Topic  5 : 0.033*"hospital" + 0.031*"nurse" + 0.009*"pain" +
Topic  6 : 0.029*"pain" + 0.012*"treatment" + 0.011*"surgery
Topic  7 : 0.014*"pregnancy" + 0.012*"facility" + 0.012*"bab
```



Visualization of the best topic model



07

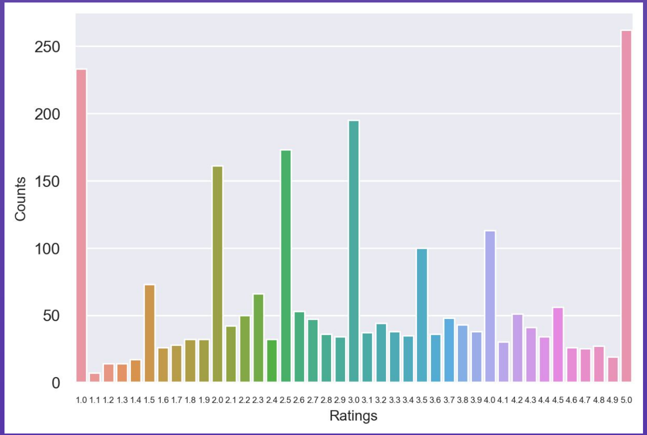
Supervised learning

Describe the plan for supervised learning
Describe all supervised learning activities
Compare the models

All supervised learning activities

- Create textual features from the cleaned review each business
- Transform the target variable
- Create training data and test data sets for full feature set
- Create training data and test data sets for SVD feature set
- Define three classifiers with hyperparameter tuning
- Feature set summary

Target variable and Train/Test Data Set




```
0    1365
1    1103
Name: Target, dtype: int64
```

Train/Test Ratio : 8:2

Stratified 10 Folding Method

Three classifiers with hyperparameter tuning

SVM	Random Forest	Gradient Boosting
'C': [0.1, 1, 10, 100], 'gamma': [1, 0.1, 0.01, 0.001], 'kernel': ['rbf', 'linear', 'poly', 'sigmoid']}	'min_samples_split': [90, 120, 150], 'n_estimators' : [100, 200, 300], 'max_depth': [3, 5, 8, 10], 'max_features':['sqrt','log2']	'learning_rate': [0.01, 0.1, 1], 'n_estimators' : [100, 200, 500,1000], 'max_features' : ['sqrt','log2']

Feature sets for supervised learning

BOW

Frequency-based
Bag-of-words

TF-IDF

Term frequency–inverse
document frequency

Glove

Glove.twitter.27B.200d
Pre-train model

Language tags

NER tags and sentiment
analysis features

Topic Vectors

Based on the best topic model

Hybrid

Glove and NER+SA

BOW

Frequency-based Bag-of-words

get	doctor	...	fast	attitude	continue	late	especially	wish	face	company	terrible	hard
1	0	...	1	0	0	0	0	0	0	0	0	0
5	0	...	0	0	0	0	0	0	0	0	0	1
10	2	...	1	1	0	0	2	0	0	0	0	0

```
300
['get', 'doctor', 'time', 'care', 'would', 'staff', 'see', 'call', 'office', 'wait', 'take', 'say', 'tell', 'make', 'patient', 'appointment', 'come', 'back', 'one'
```

TF-IDF

Term frequency–inverse document frequency

```
# create the reduced token list
# with a minimal TF - document frequency count, 2468/10=246
# with a max TF -ocument frequency count, 2468/3=822
token_list =[]
for token_id, count in dictionary.dfs.items():
        if count in range(246,1234):
            token_list.append(dictionary.get(token_id))

print(len(token_list)) # check the token list length
print(token_list[:10]) # see the list of tokens
```

```
904
['home', 'locate', 'center', 'dad', 'felt', 'longer', 'drive', 'become', 'difficult', 'physical']
```

Glove

Glove.twitter.27B.200d

Pre-train model

	indexID	AWE1	AWE2	AWE3	AWE4	AWE5	AWE6	AWE7	AWE8	AWE9	...	AWE191	AWE192	AWE193	AWE194
0	0	0.026059	0.124322	-0.059792	-0.018517	0.020472	0.085809	0.678513	-0.011019	-0.026834	...	-0.108237	0.061750	0.132979	-0.137457
1	1	-0.071928	0.044708	-0.019660	-0.026914	0.008438	0.089707	0.600606	-0.064142	-0.039016	...	-0.025482	0.034839	0.077921	-0.064481
2	2	0.035490	0.052178	-0.027157	-0.042163	0.029677	0.118611	0.701179	-0.008998	-0.045940	...	-0.072663	0.017579	0.089819	-0.096522
3	3	-0.037787	0.174142	-0.174670	0.106792	0.116708	0.038169	0.652175	-0.071135	0.142023	...	-0.110416	0.042387	0.046308	-0.206192
4	4	-0.016078	0.031949	-0.031174	0.022802	0.016231	0.078538	0.623929	-0.048225	-0.020598	...	-0.003541	-0.004253	0.073177	-0.066532

5 rows × 201 columns

Topic Vectors

Based on the best topic model

```
[(0,
  '0.032*"son" + 0.029*"child" + 0.025*"daughter" + 0.022*"kid" + 0.014*"love" + 0.014*"pediatrician" + 0.013*"baby" + 0.007*"sick" + 0.007*"nurse" + 0.007*"little"
+ 0.007*"old" + 0.006*"parent" + 0.006*"bring" + 0.006*"amazing" + 0.005*"shot" + 0.005*"question" + 0.005*"concern" + 0.005*"happy" + 0.005*"year_old" +
0.004*"front"'),
 (1,
  '0.007*"pay" + 0.006*"send" + 0.005*"speak" + 0.005*"receptionist" + 0.005*"receive" + 0.005*"referral" + 0.005*"finally" + 0.004*"horrible" + 0.004*"follow" +
0.004*"front" + 0.004*"unprofessional" + 0.004*"bill" + 0.004*"return" + 0.004*"result" + 0.004*"today" + 0.004*"show" + 0.004*"someone" + 0.004*"anything" +
0.003*"answer" + 0.003*"front_desk"'),
 (2,
  '0.023*"urgent_care" + 0.008*"clinic" + 0.008*"nurse" + 0.008*"test" + 0.007*"facility" + 0.006*"front_desk" + 0.005*"clean" + 0.005*"location" + 0.005*"today" +
0.005*"pain" + 0.005*"pay" + 0.005*"send" + 0.005*"son" + 0.004*"receptionist" + 0.004*"daughter" + 0.004*"prescription" + 0.004*"sick" + 0.004*"front" +
0.004*"ray" + 0.004*"horrible"'),
 (3,
  '0.012*"family" + 0.011*"health" + 0.011*"listen" + 0.009*"highly_recommend" + 0.008*"physician" + 0.008*"love" + 0.008*"concern" + 0.007*"practice" +
0.007*"husband" + 0.007*"thorough" + 0.007*"happy" + 0.007*"medication" + 0.006*"knowledgeable" + 0.006*"wonderful" + 0.006*"question" + 0.006*"helpful" +
0.006*"truly" + 0.005*"amazing" + 0.005*"test" + 0.005*"make_feel"'),
 (4,
  '0.023*"surgery" + 0.018*"procedure" + 0.011*"eye" + 0.009*"result" + 0.008*"treatment" + 0.008*"amazing" + 0.008*"make_feel" + 0.008*"skin" +
0.008*"highly_recommend" + 0.007*"love" + 0.007*"comfortable" + 0.007*"happy" + 0.006*"consultation" + 0.006*"everyone" + 0.005*"amaze" + 0.005*"everything" +
0.005*"face" + 0.005*"vision" + 0.004*"answer_question" + 0.004*"nurse"'),
 (5,
  '0.035*"hospital" + 0.033*"nurse" + 0.009*"pain" + 0.008*"kaiser" + 0.006*"surgery" + 0.005*"baby" + 0.005*"husband" + 0.004*"home" + 0.004*"mom" + 0.004*"son" +
0.004*"put" + 0.004*"stay" + 0.004*"finally" + 0.004*"horrible" + 0.004*"emergency_room" + 0.004*"bed" + 0.004*"admit" + 0.004*"name" + 0.004*"around" +
0.003*"sit"'),
 (6,
  '0.032*"pain" + 0.013*"treatment" + 0.011*"foot" + 0.011*"surgery" + 0.007*"highly_recommend" + 0.006*"physical_therapy" + 0.006*"massage" + 0.006*"life" +
0.006*"injury" + 0.006*"start" + 0.006*"knee" + 0.005*"chiropractor" + 0.005*"therapy" + 0.005*"body" + 0.005*"session" + 0.005*"amazing" + 0.004*"felt" +
0.004*"everyone" + 0.004*"knowledgeable" + 0.004*"neck"'),
 (7,
  '0.014*"pregnancy" + 0.012*"facility" + 0.012*"baby" + 0.012*"nurse" + 0.007*"ultrasound" + 0.007*"pregnant" + 0.006*"woman" + 0.006*"mom" + 0.006*"hospital" +
0.006*"love" + 0.005*"deliver" + 0.004*"home" + 0.004*"family" + 0.004*"send" + 0.004*"mother" + 0.004*"everything" + 0.004*"speak" + 0.004*"stay" +
0.004*"horrible" + 0.004*"let"')]
```

TV1	TV2	TV3	TV4	TV5	TV6	TV7	TV8
0.000382	0.001439	0.000627	0.292977	0.000448	0.000333	0.398625	0.305168
0.000228	0.000867	0.196523	0.000449	0.511811	0.196848	0.000239	0.093035
0.000332	0.001259	0.229730	0.100876	0.458956	0.000289	0.000349	0.208209

Language tags

NER tags and sentiment analysis features

ORG	NER_DATE	NER_PERSON	NER_MONEY	polarity	subjectivity	NLTK_Compound	NER_ORG_Scaled	NER_DATE_Scaled	NER_PERSON_Scaled	NER_MONEY_Scaled
0	3	0	0	0.405979	0.642869	0.962729	-0.429661	-0.497454	-0.381592	-0.257784
2	5	0	0	0.391728	0.585269	0.799167	0.204741	-0.356814	-0.381592	-0.257784
0	6	0	0	0.330723	0.580852	0.664585	-0.429661	-0.286494	-0.381592	-0.257784
0	0	0	0	-0.300000	0.600000	-0.440400	-0.429661	-0.708414	-0.381592	-0.257784
1	3	0	0	0.225908	0.596353	0.538986	-0.112460	-0.497454	-0.381592	-0.257784

Hybrid

Glove and NER+SA

...	AWE198	AWE199	AWE200	polarity	subjectivity	NLTK_Compound	NER_ORG_Scaled	NER_DATE_Scaled	NER_PERSON_Scaled	NER_MONEY_Scaled
...	0.053482	-0.005715	-0.064971	0.405979	0.642869	0.962729	-0.429661	-0.497454	-0.381592	-0.257784
...	0.074525	0.041948	-0.050511	0.391728	0.585269	0.799167	0.204741	-0.356814	-0.381592	-0.257784
...	0.070167	0.056712	0.003231	0.330723	0.580852	0.664585	-0.429661	-0.286494	-0.381592	-0.257784

Feature set summary

Type	Model	Accuracy	Recall	Precision	AUC
Data-driven	SVM with full bow feature set	0.866397	0.855204	0.847534	0.851351
	RF with full bow feature set	0.834008	0.733032	0.875676	0.798030
	GB with full bow feature set	0.860324	0.809955	0.868932	0.838407
	SVM with SVD bow feature set	0.823887	0.778281	0.819048	0.798144
	RF with SVD bow feature set	0.787449	0.687783	0.808511	0.743276
	GB with SVD bow feature set	0.834008	0.800905	0.823256	0.811927
	SVM with full tfidf feature set	0.872470	0.828054	0.879808	0.853147
	RF with full tfidf feature set	0.850202	0.751131	0.897297	0.817734
	GB with full tfidf feature set	0.870445	0.814480	0.886700	0.849057
	SVM with SVD tfidf feature set	0.862348	0.805430	0.876847	0.839623
	RF with SVD tfidf feature set	0.827935	0.746606	0.850515	0.795181
	GB with SVD tfidf feature set	0.858300	0.814480	0.861244	0.837209
	SVM with full glove feature set	0.854251	0.800905	0.863415	0.830986
	RF with full glove feature set	0.827935	0.746606	0.850515	0.795181
	GB with full glove feature set	0.868421	0.850679	0.854545	0.852608
	SVM with SVD glove feature set	0.864372	0.828054	0.863208	0.845266
	RF with SVD glove feature set	0.842105	0.805430	0.835681	0.820276
	GB with SVD glove feature set	0.874494	0.841629	0.873239	0.857143
Knowledge-driven	SVM with full NER and SA feature set	0.896761	0.882353	0.886364	0.884354
	RF with full NER and SA feature set	0.900810	0.877828	0.898148	0.887872
	GB with full NER and SA feature set	0.894737	0.850679	0.908213	0.878505
	SVM with SVD NER and SA feature set	0.894737	0.891403	0.875556	0.883408
	RF with SVD NER and SA feature set	0.892713	0.882353	0.878378	0.880361
	GB with SVD NER and SA feature set	0.888664	0.886878	0.867257	0.876957
	SVM with full tv feature set	0.811741	0.742081	0.820000	0.805107
	RF with full tv feature set	0.797571	0.723982	0.804020	0.790562
	GB with full tv feature set	0.815789	0.769231	0.809524	0.811355
Hybrid	SVM with full hybrid feature set	0.906883	0.886878	0.903226	0.894977
	RF with full hybrid feature set	0.888664	0.841629	0.902913	0.871194
	GB with full hybrid feature set	0.910931	0.904977	0.896861	0.900901
	SVM with SVD hybrid feature set	0.882591	0.868778	0.868778	0.868778
	RF with SVD hybrid feature set	0.854251	0.819005	0.849765	0.834101
	GB with SVD hybrid feature set	0.882591	0.859729	0.875576	0.867580

Based on its accuracy score and AUC score, among 33 models,
Gradient Boosting with full hybrid feature set is the best model here

08

Deployment Plan

Steps of Deployment Plan

Deployment Plan

Considering a proper Infrastructure for storing and processing the data.

1. Developing the model.
2. Iteration of Optimizing and testing code
3. Constant monitoring and maintenance
4. Transforming the model into a well-engineered product such as API or a website.

Thank you!

Let me know if you have question!