Honest DOC : Sentinel Analysis

Chaiyawat Suppasilp
ID: 6238023 Data Science for Healthcare Programe
Clinical epidemiology and Biostatistic Department
Ramathibodi Hospital Mahidol University

May 2020

Contents

DATA VISUALIZATIONT

- Web Scraping
- Original Data
- Data Preparation
- Data Visualization
- Choosing Hospital for Further analysis

TOPIC ANALYSIS

- NLP
- Bag of Words
- N-grams Model
- LDA Model
- Word Cloud
- Radar Chart

SENTINEL ANALYSIS

- Model selection
- Hyperparameter tuning
- Model comparison

DISCUSSION

- Topic Analysis
- Sentinel Analysis
- Analysis of Error
- Limitation
- Suggestion

DISCUSSION WILL BE ALSO INCLUDED IN EACH TOPIC

Contents

DATA VISUALIZATIONT

- Web Scraping
- Original Data
- Data Preparation
- Data Visualization
- Choosing Hospital for Further analysis

TOPIC ANALYSIS

- NLP
- Bag of Words
- N-grams Model
- LDA Model
- Word Cloud
- Radar Chart

SENTINEL ANALYSIS

- Model selection
- Hyperparameter tuning
- Model comparison

DISCUSSION

- Topic Analysis
- Sentinel Analysis
- Analysis of Error
- Limitation
- Suggestion

Web scraping

- From https://www.honestdocs.co/
- Mainly using source code from Dr. Ratchainan's Class
- All the data is in Thai languages and be translated into English by using Translator from googletrans
- The NaN, and incomplete data were manipulated accordingly.
  (including while the processing, see the code for details)
- For details, Please see in web_scraping.py

```python
#translation
from googletrans import Translator

## Translate from Thais to English
def th2en(comment):
  return Translator().translate(comment, src="th", dest="en").text
```

Original Dataset: full_data

full_data - DataFrame

Index	comment	score	hospital	en
0	โดยรวมถือว่าดีนะครับ บริการเยี่ยมพยาบาลก็ดูแลดี โดยรวมโรงพยาบาลถือว่ามีความสะ...	5	โรงพยาบาลเวิลด์เมดิคอลเซ็นเตอร์	Overall, it is considered good. Great service, nurses were well taken care of Overall, the hospital is considered clean and clean. For tho…
1	ดูแลดีมาก ดำเนินการรวดเร็ว ค่าใช้จ่ายที่พอประมาณ หมอ พยาบาลให้การดูแล ต้อน...	5	โรงพยาบาลเวิลด์เมดิคอลเซ็นเตอร์	Looks very good. Fast action. The cost is reasonable. The doctor and nurse provide good care and good care. The cost is reasonable. The d…
2	เคยเข้าไปรักษาที่โรงยาบาลนี้หลายครั้งมากๆคะ ได้เป็นทั้งผู้ป่วยนอกและผู้ป่วยในเลยคะ ...	5	โรงพยาบาลเวิลด์เมดิคอลเซ็นเตอร์	I've been to this hospital many times. Can be both an outpatient and inpatient. Nurse, nurse assistant and doctor and hospital staff.…
3	ตั้งใจคลอดธรรมชาติ และซื้อแพคเกจคลอดธรรมชาติไว้ที่นี่ ปรากฏว่าวันเจ็บท้องคลอดป...	5	โรงพยาบาลเวิลด์เมดิคอลเซ็นเตอร์	Intends to give birth naturally And buy a natural birth package here It appears that the days of labor in the uterus are not open. Until …
4	ดูแลดีตั้งแต่เดินเข้าประตูมีพยาบาลมาถามว่าเป็นอะไร มีบัตรรพ.มั้ย มีประกันรึเปล่า พู...	1	โรงพยาบาลเวิลด์เมดิคอลเซ็นเตอร์	Taking good care since walking into the door, there was a nurse asking what was wrong. Do you have a hospital card? Do you have insu…
5	บัณฑิตา ประวาลพฤกษ์	3	โรงพยาบาลเวิลด์เมดิคอลเซ็นเตอร์	Banthita Prawanphruk
6	บริการดีมาก รอไม่นาน พอแจ้งเวชทะเบียนปุ๊ปมีคนนำขึ้นไป ถือว่าการบริการพร้อมมา...	4	โรงพยาบาลเวิลด์เมดิคอลเซ็นเตอร์	Very good service, waiting for not long enough to inform the registered doctor. Considered the service is very ready. Nurse servi…
7	สถานที่สะอาด กว้างขวาง สะดวก สบาย ที่จอดรถมีพอ ราคาไม่แพงมาก จนท ...	5	โรงพยาบาลเวิลด์เมดิคอลเซ็นเตอร์	The place is clean, spacious, convenient, parking is enough, the price is not expensive, the service is good.
8	พนักงานบริการดีมากๆค่ะ ใส่ใจทุกรายละเอียด ไม่ว่าจะเป็นการตรวจ จ่ายตัง	5	โรงพยาบาลเวิลด์เมดิคอลเซ็นเตอร์	The service staff is very good. Pay attention to every detail. Whether the payment
9	โรงพยาบาลต้อนรับและดูแลเราตั้งแต่ก้าวขาลงรถ ตลอดจนการประสานงานและสอบถาม...	4	โรงพยาบาลเวิลด์เมดิคอลเซ็นเตอร์	The hospital welcomed and looked after us since stepping down the car. As well as coordinating and inquiring about symptoms until you …
10	เป็นโรงพยาบาลที่ดี มีการดูแลเอาใจใส่คนไข้ เป็นกันเองกับคนไข้มาก ค่ารักษาพยาบาล...	5	โรงพยาบาลธัญบุรี	Is a good hospital Patient care Very friendly to patients Medical expenses are not expensive, so glad to come to this hospital. The pl…
11	ให้บริการ รวดเร็ว รอคิวไม่นาน.. เจ้าหน้าที่และคุณหมอเป็นกันเอง ไปตรวจ เจ้าห...	4	โรงพยาบาลธัญบุรี	Providing fast service, waiting in line no longer .. Staff and doctors are friendly, go to check the staff and the hospital to prov…
12	บริการดีหมอและพยาบาลดูแลดีพูดจาดีค่ะ เป็นกันเองรอคิวไม่นาน. ติดตามอาการพนัก...	5	โรงพยาบาลธัญบุรี	Good service, doctors and nurses, good care, speak well. Friendly, waiting in line not long Follow the staff who work well. Give us use…

Data Visualization

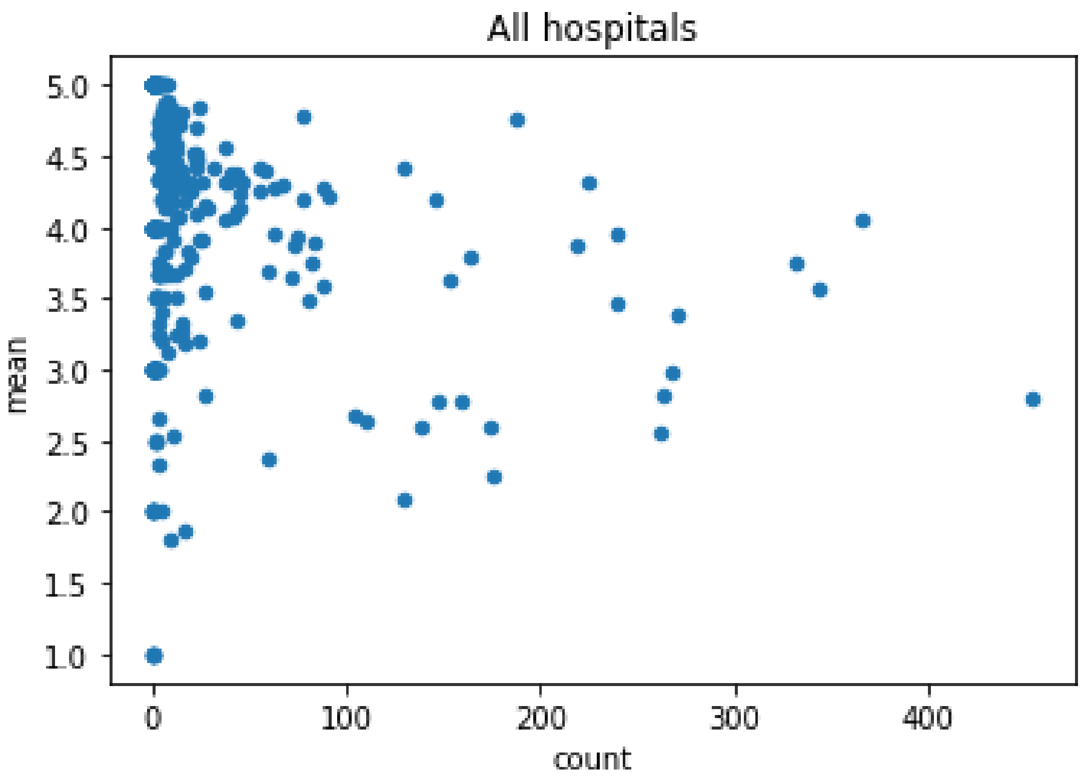
- I created a new data frame to demonstrate the number of comments (variable: 'count') and the mean of 'score' of each hospital
- The overall mean = 3.63 (S.D.=1.55)

Visual_df - DataFrame

Index	mean	count
Wannisa Kongmee	4.375	8
คลินิกศูนย์แพทย์พัฒนา	4.46154	13
มิตรไมตรีคลินิกเวชกรรม สาขาบึงคำพร้อย	4.25	8
มิตรไมตรีคลินิกเวชกรรม สาขาพรประภานิมิตร	4.8	15
มิตรไมตรีคลินิกเวชกรรม สาขาลาซาล	5	1
มิตรไมตรีคลินิกเวชกรรม สาขาวัดพระเงิน	4.78205	78
มิตรไมตรีคลินิกเวชกรรม สาขาเครือสหพัฒน์	4.5	2
มิตรไมตรีคลินิกเวชกรรม สาขาเทพประสิทธิ์	5	1
ศูนย์การแพทย์กาญจนาภิเษก	4	1
ศูนย์การแพทย์สมเด็จพระเทพรัตน์	4.6	5
ศูนย์ศรีพัฒน์	4.33333	3
ศูนย์เวชศาสตร์ผู้สูงอายุ มหาวิทยาลัยเชียงใหม่	4	1
สถาบันประสาทวิทยา	4.33333	3
สถาบันมะเร็งแห่งชาติ	4.83333	6

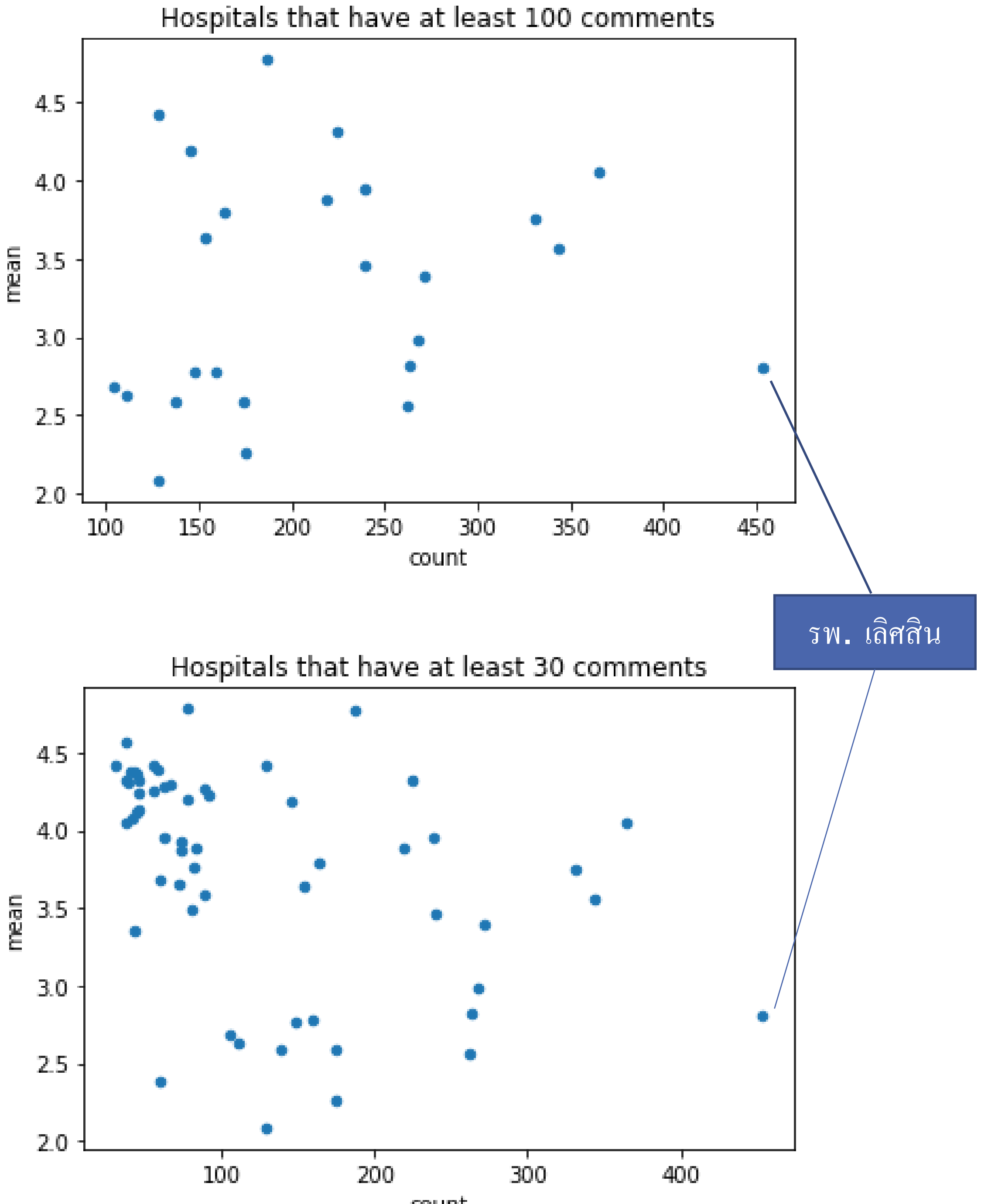
Data Visualization

- From count-mean plot of All hospitals, you may see that there are many hospitals that have low counts
- From my opinion, low-count data may not be a good candidate for further analyses.
  - The data is not a representative of the users
  - The data may have high degree of bias
  - Too small data set for machine learning



Data Visualization

- I decided to visualized only hospitals that have counts at least 30 comments and 100 comments.
- The outlier with about 450 comments and score of 2.8 is Lerdsin hospital. Because of these sufficient number of comments, this hospital should go to see the comment with some analyses e.g. NLP to identify the problems. (unfortunately, I decided not to choose this hospital, due to limited of my time.)



Data Visualization

- TOP 10 and Bottom 10 That have counts at least 30 counts

count_30 - DataFrame

Index	mean	count
มิตรไมตรีคลินิกเวชกรรม สาขาวัดพระเงิน	4.78205	78
โรงพยาบาลจุฬาภรณ์	4.77005	187
โรงพยาบาลนครธน	4.56757	37
โรงพยาบาลสมิติเวช ธนบุรี	4.41935	31
โรงพยาบาลศิริราช	4.4186	129
โรงพยาบาลวชิรพยาบาล	4.41818	55
โรงพยาบาลนวมินทร์ 1	4.39655	58
โรงพยาบาลบีเอ็นเอช	4.375	40
โรงพยาบาลเซนต์หลุยส์	4.37209	43
โรงพยาบาลลานนา	4.36364	44

count_30 - DataFrame

Index	mean	count
โรงพยาบาลนพรัตนราชธานี	2.08527	129
โรงพยาบาลเกษมราษฎร์ บางแค	2.25714	175
โรงพยาบาลตากสิน	2.38333	60
โรงพยาบาลศิครินทร์	2.56489	262
โรงพยาบาลเกษมราษฎร์ ประชาชื่น	2.58696	138
โรงพยาบาลวิภาราม	2.59195	174
โรงพยาบาลบางปะกอก 3	2.63063	111
โรงพยาบาลเจริญกรุงประชารักษ์	2.68571	105
โรงพยาบาลราษฎร์บูรณะ	2.77027	148
โรงพยาบาลราชวิถี	2.77987	159

Data Visualization

- TOP 10 and Bottom 10 That have counts at least 100 counts

count_100 - DataFrame

Index	mean	count
โรงพยาบาลจุฬาภรณ์	4.77005	187
โรงพยาบาลศิริราช	4.4186	129
โรงพยาบาลบำรุงราษฎร์	4.31556	225
โรงพยาบาลเวชธานี	4.19178	146
โรงพยาบาลศิริราช ปิยมหาราชการุณย์	4.04932	365
โรงพยาบาลกรุงเทพ	3.94979	239
โรงพยาบาลรามาธิบดี	3.88128	219
โรงพยาบาลพญาไท 1	3.79268	164
โรงพยาบาลจุฬาลงกรณ์	3.74924	331
โรงพยาบาลพญาไท 3	3.63636	154

count_100 - DataFrame

Index	mean	count
โรงพยาบาลนพรัตนราชธานี	2.08527	129
โรงพยาบาลเกษมราษฎร์ บางแค	2.25714	175
โรงพยาบาลศิครินทร์	2.56489	262
โรงพยาบาลเกษมราษฎร์ ประชาชื่น	2.58696	138
โรงพยาบาลวิภาราม	2.59195	174
โรงพยาบาลบางปะกอก 3	2.63063	111
โรงพยาบาลเจริญกรุงประชารักษ์	2.68571	105
โรงพยาบาลราษฎร์บูรณะ	2.77027	148
โรงพยาบาลราชวิถี	2.77987	159
โรงพยาบาลเลิดสิน	2.80132	453

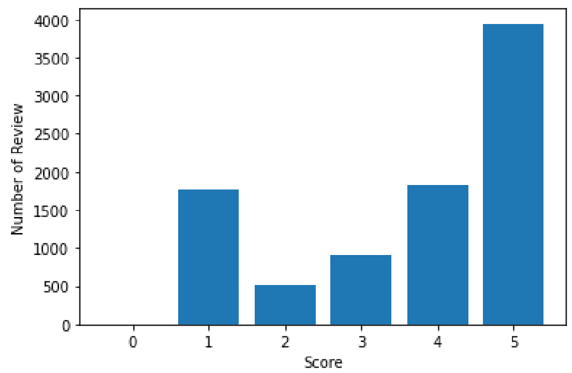
Sentiment polarity

- Positive if the number of stars is greater than 3 (or 4,5) (variable sentiment = 1)
- Neutral if the number of stars is equal to 3 (will be excluded for sentiment analysis)
- Negative if the number of stars is less than 3 (or 1,2) (variable sentiment = 0)

```
# convert to sentiment type
# less than 3 is a negative sentiment = 0
df.loc[df['score'] < 3, 'sentiment'] = 0
# equal to 3 is a neutral sentiment, this group will not be used to build the model
df.loc[df['score'] == 3, 'sentiment'] = 'nan'
# more than 3 is a positive sentiment = 1
df.loc[df['score'] > 3, 'sentiment'] = 1
```

Sentiment polarity

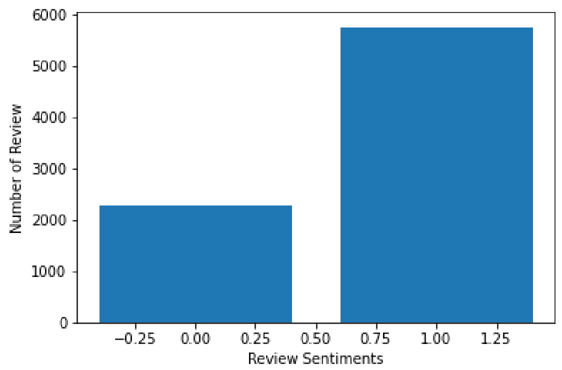
```
In [70]: score_graph(full_data)
```




```
        comment  percentage
score
0             2    0.022369
1          1762   19.706968
2           510    5.704060
3           909   10.166648
4          1816   20.310927
5          3942   44.089028
```

Score 0 are not included for further analyses.

```
In [80]: sentiment_graph(df_full)
```




```
           comment  percentage
sentiment
0.0           2274   28.311753
1.0           5758   71.688247
```

df_full - DataFrame

Index	comment	score	hospital	en	sentiment
8422	พอดีดิฉันตกเลือด แล้วทาง...	2	Wannisa Kongmee	I was bleeding. And…	0
8423	แม่ให้นมเป็นมาเป็นอาทิตย์...	4	Wannisa Kongmee	Mother is breastfeeding…	1
8424	คุณหมอและพยาบาลดูแลดี...	5	Wannisa Kongmee	The doctors and nurses ar…	1
8425	รพนี้ให้บริการดีดูแลผู้ป่วยไ...	5	Wannisa Kongmee	This hospital provides good…	1
8426	พยาบาลและหมอสูตินารีบริ...	5	Wannisa Kongmee	Nurses and doctors, obst…	1
8427	คุณหมอ พยาบาล และผู้...	5	Wannisa Kongmee	The doctors, nurses and as…	1
8428	ประสบการณ์จากการคลอด...	5	Wannisa Kongmee	The experience of giving bir…	1
8429	เพิ่งไปหาหมอที่ รพ. บา...	4	Wannisa Kongmee	Just went to the doctor at…	1
3175	ตอนพบหมอ คุณหมอพูดจ...	4	คลินิกศูนย์แพทย์พัฒนา	When seeing a doctor, the d…	1
3176	ไปครั้งแรก เป็นตากุ้งยิงค่ะ...	5	คลินิกศูนย์แพทย์พัฒนา	The first visit is an e…	1
3177	หมอใส่ใจผู้ป่วยดีมากค่ะ....	4	คลินิกศูนย์แพทย์พัฒนา	The doctor cares about t…	1
3178	ที่จอดรถหายาก พ...	4	คลินิกศูนย์แพทย์พัฒนา	Rare parking The staff and…	1
3179	หมอดูแลคนไข้ดีมาก	4	คลินิกศูนย์แพทย์พัฒนา	The doctor takes good ca…	1

Choosing Hospital for Further analysis

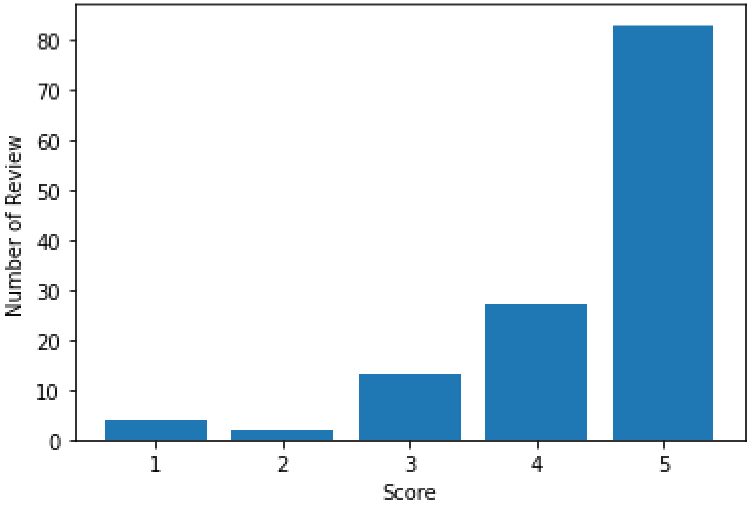
- The must is Ramathibodi Hospital (where we are studying in)
- So, the comparators should be Big medical school hospital in Thailand. Thus, I decided to choose Siriraj and Chulalongkorn Hospitals.
- Further, Rama and Siriraj have kind of private hospital running by medical-school staffs (Somdech Phra Debaratana Medical Centre (SDMC) for Rama, and Siriraj Piyamaharajkarun Hospital (SiPH) for Siriraj). I want to compare these two hospitals with the most famous private hospital in Thailand*, Bumrungrad International Hospital (BHH). But, SDMC has only 6 comments that is not insufficient. So, I will compare only SiPH and BHH.
- Next several sides will show the overall visualization of these selected hospitals.

*ref: https://healthmeth.wordpress.com/2019/08/30/5-top-best-private-hospital-thailand-2020

Siriraj

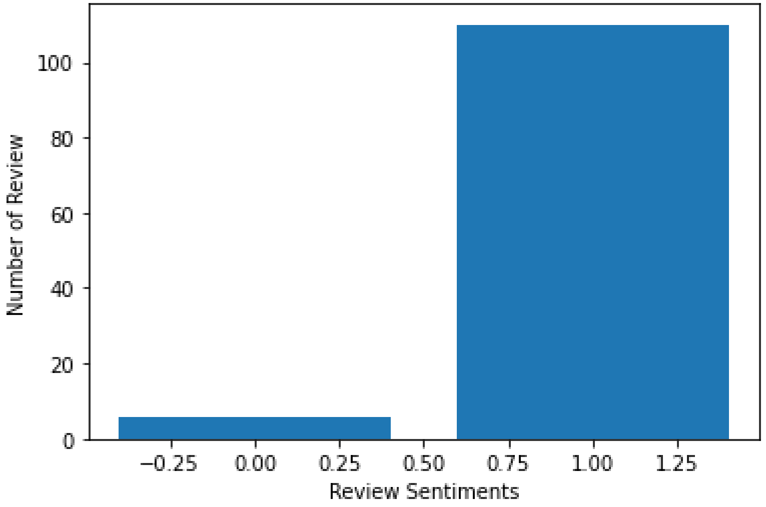
```
In [45]: siriraj_v.describe()
Out[45]:
            score
count  129.000000
mean     4.418605
std      0.957680
min      1.000000
25%      4.000000
50%      5.000000
75%      5.000000
max      5.000000
```

```
In [71]: score_graph(siriraj_v)
```




```
       comment  percentage
score
1            4    3.100775
2            2    1.550388
3           13   10.077519
4           27   20.930233
5           83   64.341085
```

```
In [81]: sentiment_graph(siriraj_df)
```

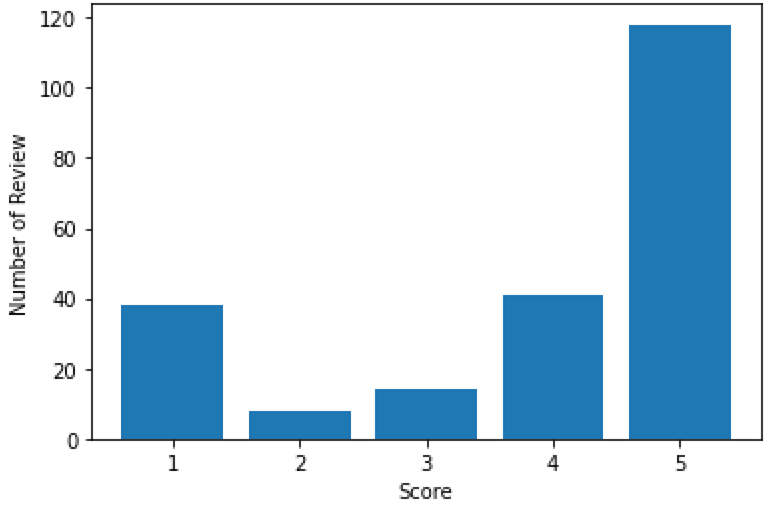



```
           comment  percentage
sentiment
0.0              6    5.172414
1.0            110   94.827586
```

Ramathibodi

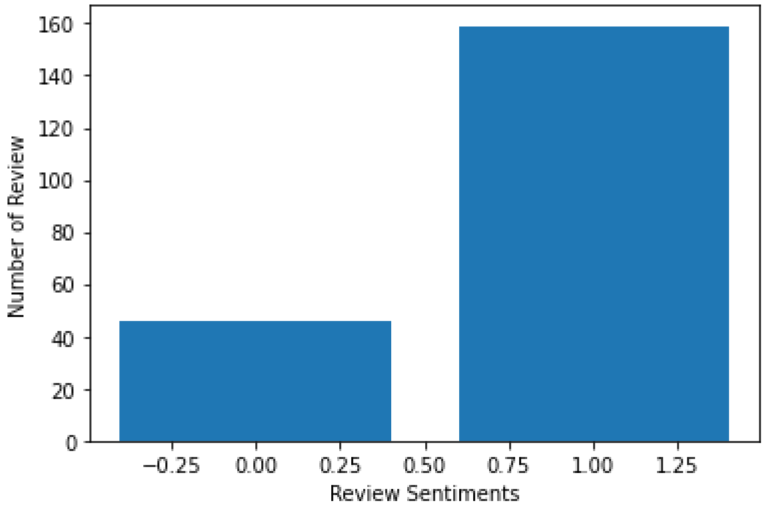
```
In [48]: rama_v.describe()
Out[48]:
            score
count  219.000000
mean     3.881279
std      1.518862
min      1.000000
25%      3.000000
50%      5.000000
75%      5.000000
max      5.000000
```

```
In [72]: score_graph(rama_v)
```



score	comment	percentage
1	38	17.351598
2	8	3.652968
3	14	6.392694
4	41	18.721461
5	118	53.881279

```
In [79]: sentiment_graph(rama_df)
```



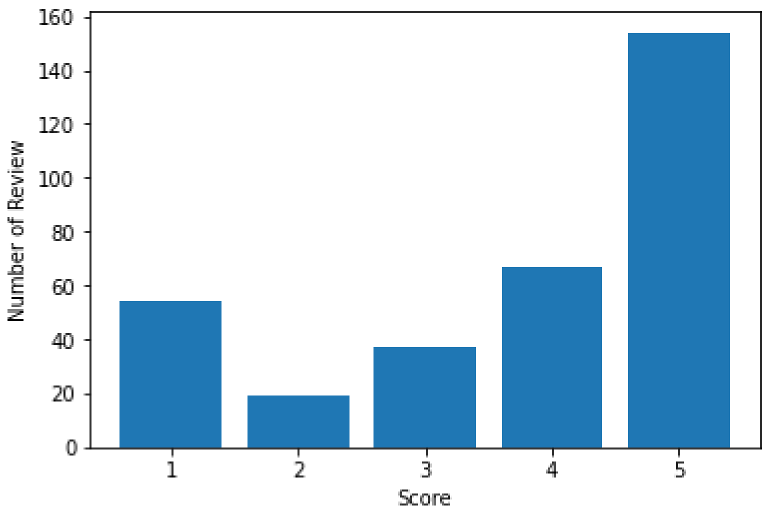
sentiment	comment	percentage
0.0	46	22.439024
1.0	159	77.560976

Chulalongkorn

```
In [50]: chula_v.describe()
Out[50]:
```

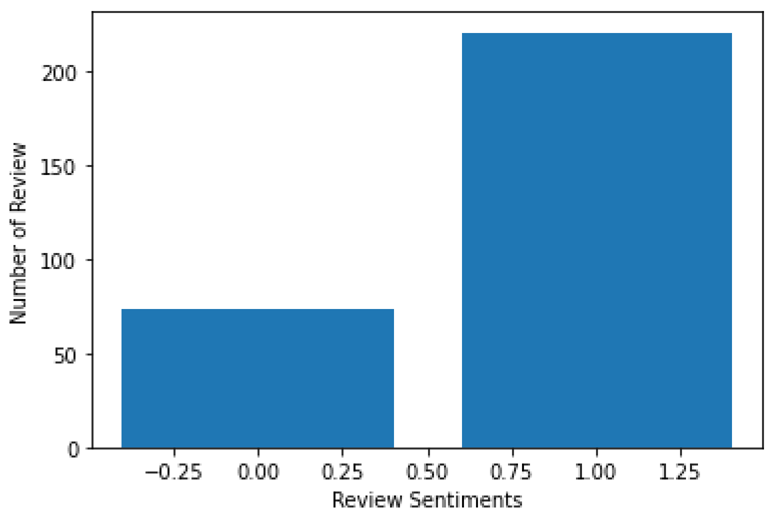
	score
count	331.000000
mean	3.749245
std	1.489547
min	1.000000
25%	3.000000
50%	4.000000
75%	5.000000
max	5.000000

```
In [73]: score_graph(chula_v)
```



	comment	percentage
score		
1	54	16.314199
2	19	5.740181
3	37	11.178248
4	67	20.241692
5	154	46.525680

```
In [82]: sentiment_graph(chula_df)
```



	comment	percentage
sentiment		
0.0	73	24.829932
1.0	221	75.170068

SI vs RA vs CU

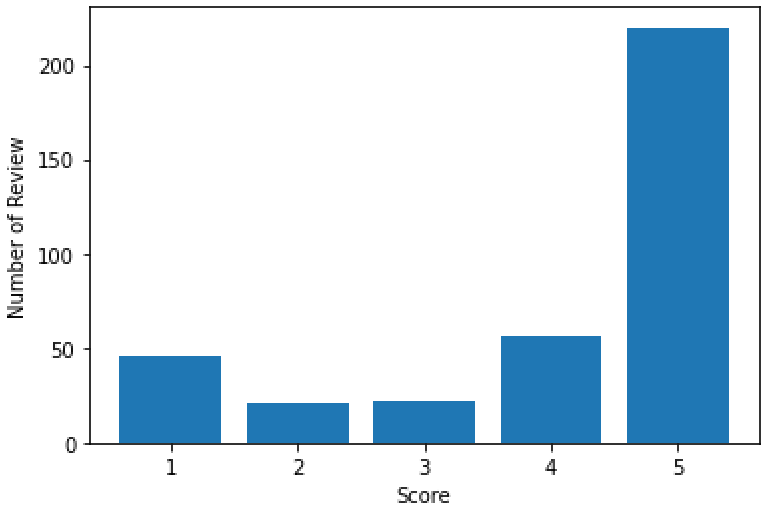
- Siriraj has the highest score of 4.42 (SD= 0.96), which is significantly higher mean as compared to the other two.
- Ramathibodi and Chulalongkorn hospitals have score of 3.88 (1.52) and 3.75 (1.49), respectively.
- By score graph, you may see that the distribution of Rama and Chula are quite the same, while the Siriraj has no peak at score 1.

Comparison	Independent t-test	P-value
Siriraj VS Rama	3.6299	0.0003
Siriraj VS Chula	4.7367	0.0001
Rama VS Chula	0.9936	0.3208

Siriraj Piyamaharajkarun Hospital (SiPH)

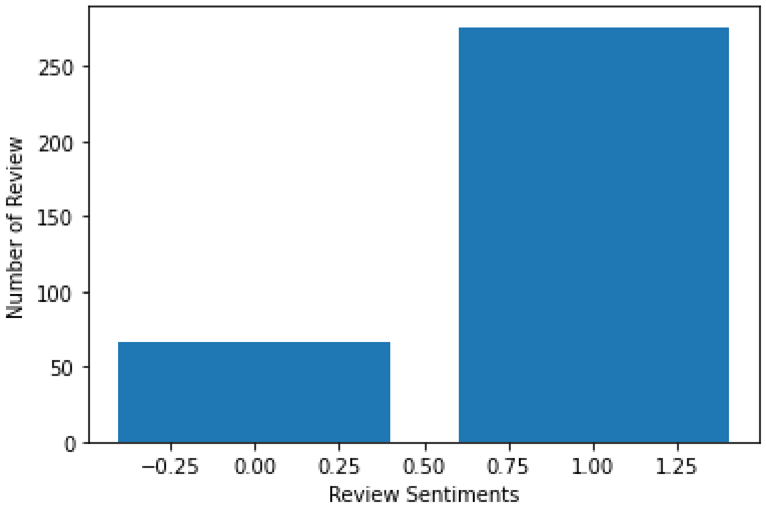
```
In [53]: siph_v.describe()
Out[53]:
            score
count  365.000000
mean     4.049315
std      1.424966
min      1.000000
25%      4.000000
50%      5.000000
75%      5.000000
max      5.000000
```

```
In [74]: score_graph(siph_v)
```




```
       comment  percentage
score
1           46   12.602740
2           21    5.753425
3           22    6.027397
4           56   15.342466
5          220   60.273973
```

```
In [83]: sentiment_graph(siph_df)
```

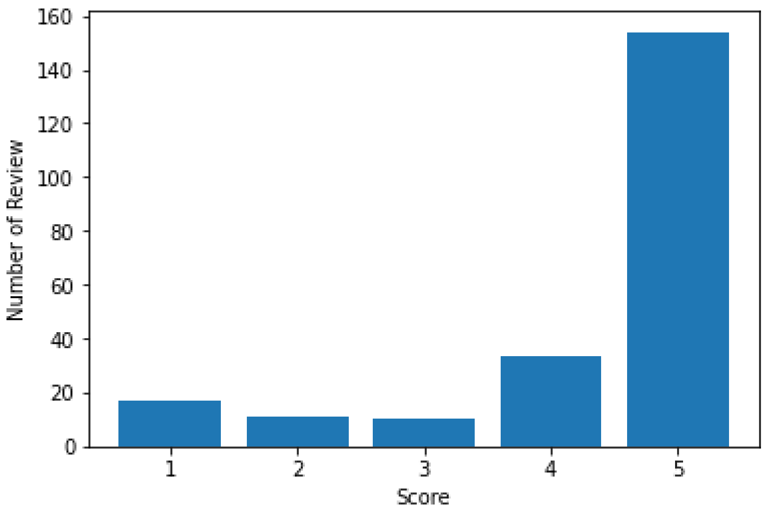



```
           comment  percentage
sentiment
0.0             67   19.533528
1.0            276   80.466472
```

Bumrungrad International Hospital(BHH)

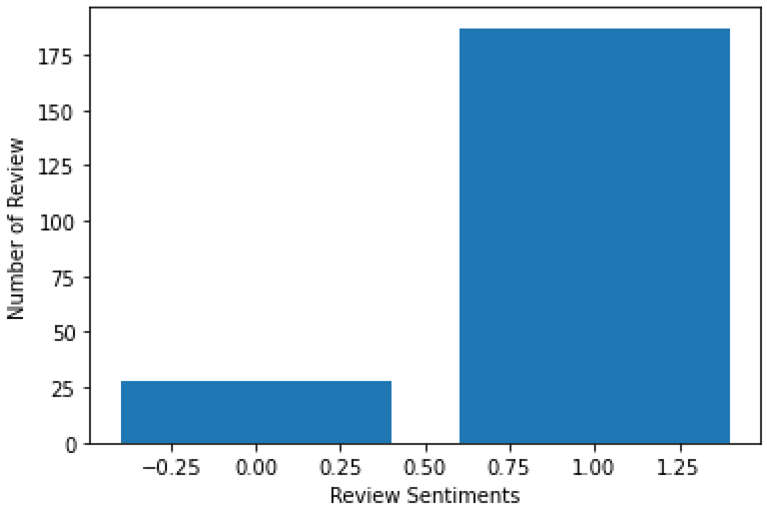
```
In [56]: bhh_v.describe()
Out[56]:
            score
count  225.000000
mean     4.315556
std      1.229466
min      1.000000
25%      4.000000
50%      5.000000
75%      5.000000
max      5.000000
```

```
In [75]: score_graph(bhh_v)
```




```
       comment  percentage
score
1           17    7.555556
2           11    4.888889
3           10    4.444444
4           33   14.666667
5          154   68.444444
```

```
In [84]: sentiment_graph(bhh_df)
```




```
           comment  percentage
sentiment
0.0             28   13.023256
1.0            187   86.976744
```

SIPH vs BHH

- These two hospitals have small difference of the score.
- For SIPH, mean is 4.05 (SD 1.42)
- For BHH, mean is 4.32 (SD 1.23)
- The distributions are quite the same

Comparison	Independent t-test	P-value
SIPH VS BHH	2.3583	0.0187

Contents

DATA VISUALIZATIONT

- Web Scraping
- Original Data
- Data Preparation
- Data Visualization
- Choosing Hospital for Further analysis

TOPIC ANALYSIS

- NLP
- Bag of Words
- N-grams Model
- LDA Model
- Word Cloud
- Radar Chart

SENTINEL ANALYSIS

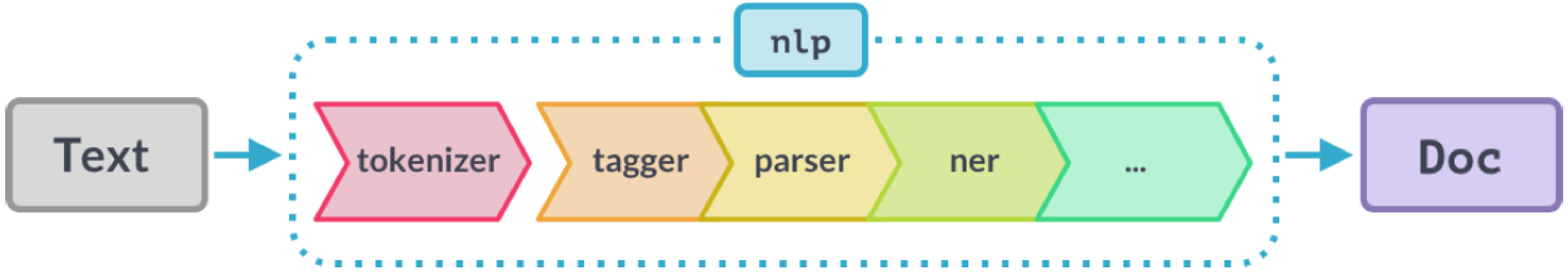
- Model selection
- Hyperparameter tuning
- Model comparison

DISCUSSION

- Topic Analysis
- Sentinel Analysis
- Analysis of Error
- Limitation
- Suggestion

Natural Languages Processing (NLP)

- In these assignment, I mainly use Spacy library with 'en_core_web_sm' statistical package for NLP task.
- The Spacy allows us to perform Tokenization, tagger, parser, name entity recognition, including stop word removing.
- The statistical package (sm = small) can act like pre-train learning that enable spaCy to predict linguistic attributes in context better. For example, part-of-speech tagging, syntactic dependencies, named entities.
- However, some tasks I will use 'en_core_web_md' (medium size) if word vectorization is needed (sm cannot perform well and lg is too large for me). And also, genism in the context of word embedded task.



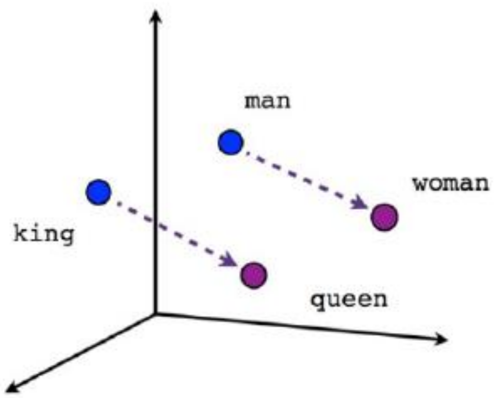
Main NLP code


```python
## NLP processing

nlp = spacy.load('en_core_web_sm')

data = df_full.copy()
data = data.reset_index()
del data['index']
L  = []
for text in data['en']:
    doc = nlp(text)
    # including Tokenization
    # including Lower case
    # Creating Lemma words for this row
    lemmas = [token.lemma_ for token in doc]
    # Creating stop-words for this row
    stopwords = spacy.lang.en.stop_words.STOP_WORDS
    # Removing non-alphabetic characters
    # Lemmatization
    # Removeing stop-words
    a_lemmas = [lemma for lemma in lemmas if lemma.isalpha() and lemma not in stopwords]
    b_lemmas = ' '.join(a_lemmas)
    L.append(b_lemmas)

df_L = pd.DataFrame(L, columns=['docc'])
data = pd.merge(data, df_L, left_index=True, right_index=True)
data.to_csv(r'D:/_RADS611 NLP/Assignment/honest_doc_full_post_nlp.csv')
```

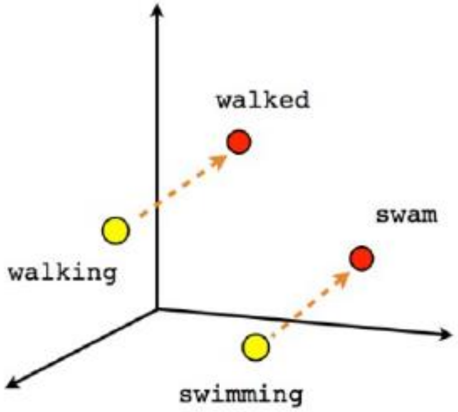
- More details about NLP, please see in Python code

Gensim

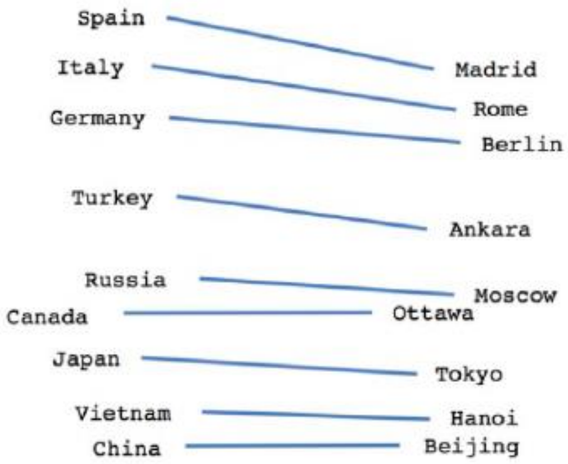
- Popular open-source NLP library
- Uses top academic models to perform complex tasks
  - Building document or word vectors
  - Performing topic identification and document comparison



Male-Female



Verb tense



Country-Capital

Post-NLP data

nlp_data - DataFrame

Index	namec	comment	score	ospita	en	ntime	docc
0	0	ดูแลดีตั้งแต่เดินเข้าประตูมีพยาบาลมาถามว่าเป็นอะไร มีบัตรรพ.มั้ย มีประกันรึเปล่า …	1	โรงพยา…	Taking good care since walking into the door, there was a nurse asking what was wrong. Do …	0	good care walk door nurse ask wrong hospital card insurance lot finish check insurance right sit w…
1	1	พยาบาลห้องคลอดดูแลและใส่ใจดีค่ะ แต่หมอจะดุๆพูดจาไม่ค่อยรักษาน้ำใจเท่าไร น่าจ…	2	โรงพยา…	Maternity room care and attention But the doctor will be fierce, speak with not much k…	0	maternity room care attention doctor fierce speak kindness change speech encourage patient instead…
2	2	มีการบริการไม่ค่อยดี พูดจาน้ำเสียงไม่ดี	2	โรงพยา…	Not very good service Speak in a bad tone	0	good service speak bad tone
3	3	พาแฟนส่งตัวมารักษาต่อที่นี่ตามสิทธิ์.. แต่เหมือนพาแฟนมานอนรอความตาย.. แ…	1	โรงพยา…	Bringing my girlfriend back to treat here as right .. But like bringing a girlfriend to s…	0	bring girlfriend treat right like bring girlfriend sleep wait death heart hurt hour wait…
4	4	ไม่ประทับใจ สถานที่แคบ ไม่มีที่เพียงพอให้คนสูงอายุหรือคนที่เจ็บหนัก ล่าช้าทุกขั้น…	1	โรงพยา…	Not impressed. Narrow location, not enough place for the elderly or seriously injured p…	0	impressed narrow location place elderly seriously injure people delay step unsuitable parking plac…
5	5	เป็น รพ. ที่ผู้ป่วย 50% เป็น 30 บาทรักษาทุกโรค อีก 47 % เป็น ปก…	2	โรงพยา…	Is a hospital in which 50% of patients are 30 baht to treat all diseases. Another 47% are …	0	hospital patient baht treat disease standard parking lot assistant manager pak lada da doctor…
6	6	ร.พ.อาคารทันสมัยเข้ากับการต้อนรับอาเซียนแต่ควรจัดระบบบริหารการทำงานเจ้าหน้า…	1	โรงพยา…	The modern hospital building meets ASEAN, but the administrative system should be organize…	0	modern hospital building meet asean administrative system organize staff new buildin…
7	7	การจัดการค่อนข้างวุ่นวาย จำนวนผู้ป่วยเยอะแต่ไม่ขยายห้องเพิ่ม บุคลาการทางการแพท…	2	โรงพยา…	Management is quite chaotic. A lot of patients but not expanding more rooms. Littl…	0	management chaotic lot patient expand room little medical personnel hospital care foreigner people…
8	8	ไปตรวจอาการผิดปรกติของตา เนื่องจากสายตาแย่ลง มองในที่มืดไม่ชัด	1	โรงพยา…	To check for abnormalities in the eye Due to worsening eyesight Looking in the dark is no…	0	check abnormality eye worsen eyesight look dark clear
9	9	หมอไม่ค่อยมาดูอาการของคนไข้ที่พักฟื้นเลยคะ มาดูแผลแค่ครั้งเดียวในวันแรกที่ผ่า แ…	2	โรงพยา…	The doctor doesn't really look at the symptoms of patients recovering. Came to see…	0	doctor look symptom patient recover come wound day surgery come leave hospital
10	10	ลูกชายเกิดอุบัติเหตุค่ะ เข้ารพตามสิทธิ์ปกส. รพ.ทำการเลือกห้องพิเศษให้เองเลยค่ะ…	1	โรงพยา…	My son had an accident. Go to the hospital as per the rights. The hospital has chosen a sp…	0	son accident hospital right hospital choose special room allow pay thousand room miscellaneo…
11	11	วันที่ 28 ธันวาคม 2559 ไปใช้บริการที่ตึก C ศูนย์มะเร็งนรีเวช และ นรีเว…	2	โรงพยา…	On December 28, 2016, went to use the service at Building C, Gynecological and Gynecologic…	0	december use service building c gynecological gynecological cancer center beautiful place impr…
12	12	รักษาแย่มากค่ะและโรงพยาบาลก็สกปรก เริ่มแรกเราเกิดอุบัติเหตุที่นิ้วมือเห็นว่าเป็นรพ.…	1	โรงพยา…	The treatment is terrible and the hospital is dirty. Initially, we had an accident in whic…	0	treatment terrible hospital dirty initially accident finger nearby hospital enter ask pressu…

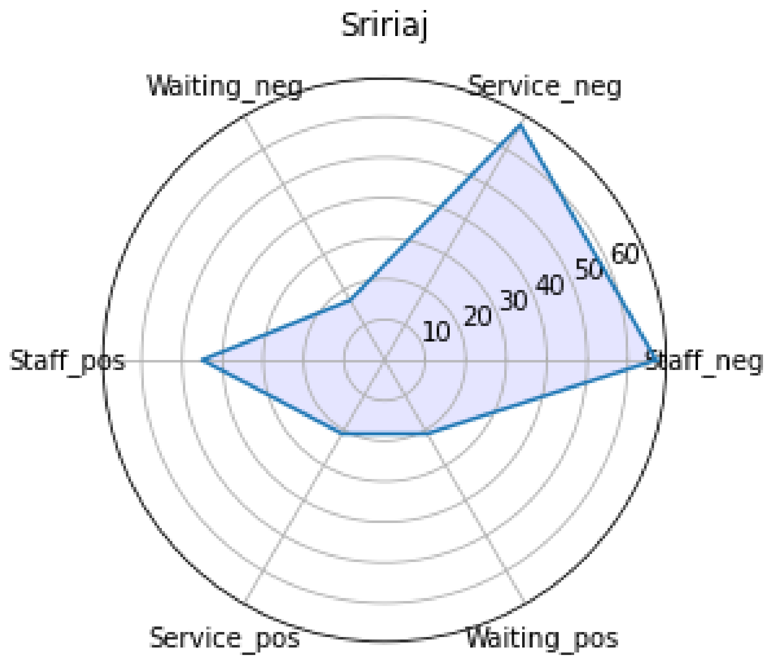
Topic Analyses

- To find the most common words that used in the comments. I will create bag of words first. And then, I will use 1-gram (=BOWs), 2-gram, 3-gram models and also term frequency–inverse document frequency (Tf-ifd) vectorization technique to convert words into vectors with Latent Dirichlet allocation LDA model to find the TOP10 word from positive sentiment group and negative sentiment group.
- Later, word clouds and radar charts will be done accordingly.

Positive group
(Sentinel=1, score=4,5)
Negative group
(Sentinel=0, score=1,2)

Radar Chart

- There are 6 axis but only 3 main topics; Staffs, Services, Waiting
- The positive and negative of the same topic are in the opposite direction for easily to be compared
- The number on the charts are percentage of the comments from the total positive/negative comments
- The number shown on the charts were partially manual manipulated due to complexity of the n-grams and meaning of the words. Further analysis should be done.



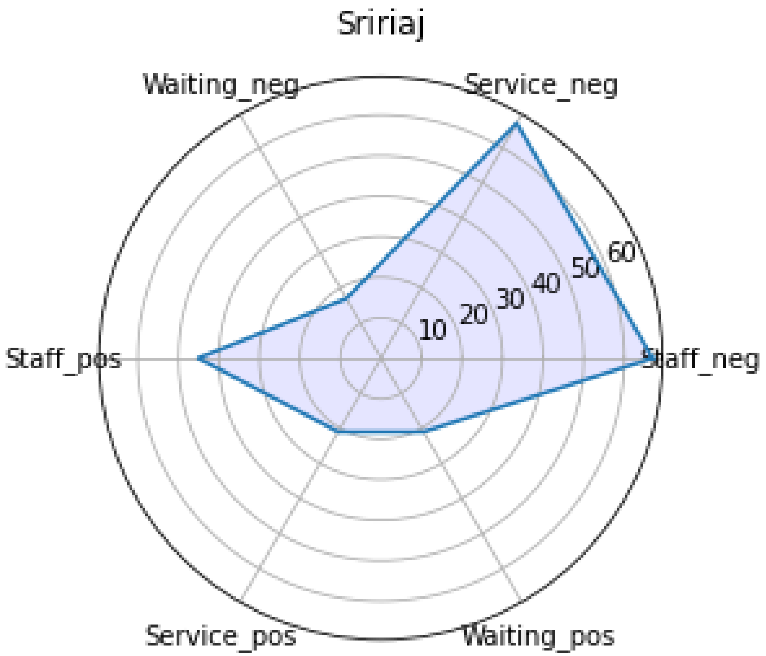
Positive Sentinel group

- Staff_pos
  comment about good staff, doctor, nurses or direct mentions about staffs in 2-gram e.g. good doctors, good nurses
- Service_pos
  comment about good services and treatments.
- Waiting_pos
  comments about waiting, long time, long queue are still occurred in positive comment

Negative Sentinel group

- Staff_neg
  comment about staff, doctor, nurses or direct mentions about staffs in 2-gram e.g. nurse speak, wait doctor
- Service_neg
  comment about good services and treatments e.g. pull body, speak badly
- Waiting_neg
  comments about waiting, long time, long queue

Siriraj



Comment:
There are 2 negative comments about treatment complications

```
Top  10 positive group

----- 10 most common 1-grams -----
doctor: 166
good: 98
hospital: 82
patient: 76
nurse: 65
time: 63
service: 52
treatment: 45
siriraj: 43
wait: 40

----- 10 most common 2-grams -----
siriraj hospital: 23
doctor nurse: 23
doctor good: 18
wait long: 13
good service: 12
service good: 11
long time: 10
good care: 9
nurse staff: 8
care patient: 7

----- 10 most common 3-grams -----
wait long time: 6
doctor nurse staff: 5
doctor good advice: 5
good advice follow: 4
advice follow treatment: 4
follow treatment doctor: 4
treatment doctor knee: 4
doctor knee joint: 4
knee joint pain: 4
wait long queue: 3
```

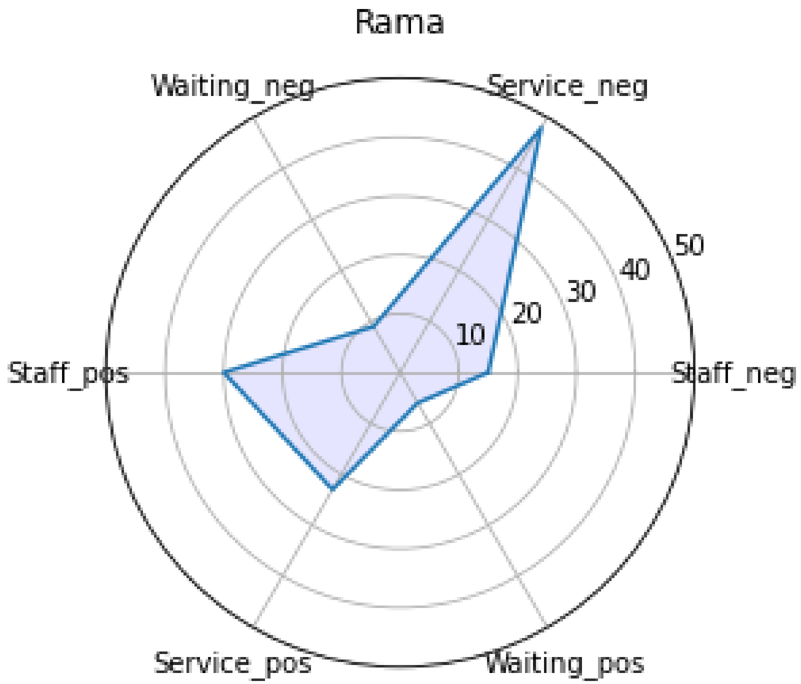
```
Top  10 negative group

----- 10 most common 1-grams -----
doctor: 11
patient: 5
good: 4
neck: 3
treatment: 3
speak: 3
surgery: 2
day: 2
time: 2
year: 2

----- 10 most common 2-grams -----
doctor enter: 1
enter fever: 1
fever surgery: 1
surgery good: 1
good body: 1
body pull: 1
pull blood: 1
blood cord: 1
cord second: 1
second round: 1

----- 10 most common 3-grams -----
doctor enter fever: 1
enter fever surgery: 1
fever surgery good: 1
surgery good body: 1
good body pull: 1
body pull blood: 1
pull blood cord: 1
blood cord second: 1
cord second round: 1
second round teacher: 1
```

Siriraj

word cloud for positive group

word cloud for negative group

Rama



Comment:
There are 3 negative words about emergency rooms.

Top 10 positive group

```
----- 10 most common 1-grams -----
doctor: 120
good: 114
hospital: 97
nurse: 74
patient: 61
service: 58
time: 42
care: 40
come: 37
thank: 37

----- 10 most common 2-grams -----
doctor nurse: 32
good service: 20
good care: 16
hospital good: 13
rama hospital: 12
ramathibodi hospital: 11
long time: 9
good good: 9
nurse good: 8
nurse speak: 8

----- 10 most common 3-grams -----
doctor nurse good: 6
nurse good care: 5
thank doctor nurse: 5
wait long time: 4
doctor nurse look: 4
doctor good care: 3
good service wait: 3
good good good: 3
doctor nurse staff: 3
nurse look patient: 3
```

Top 10 negative group

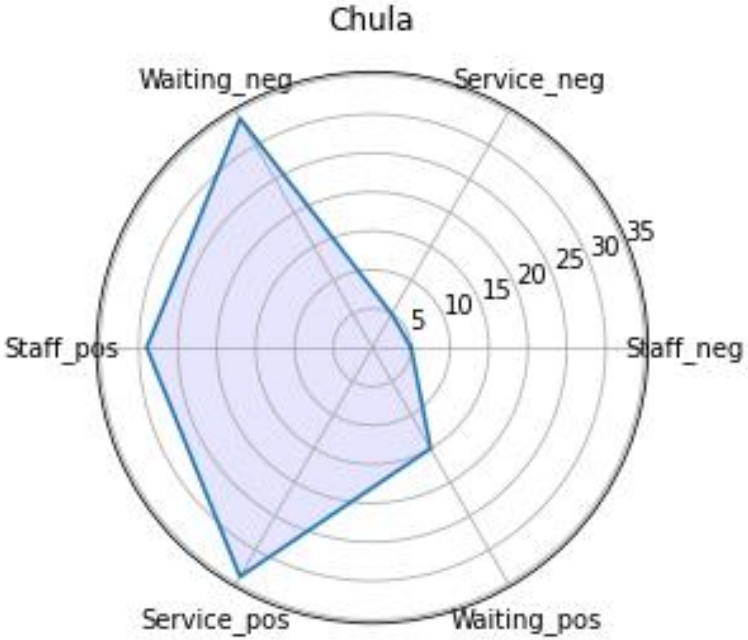
```
----- 10 most common 1-grams -----
nurse: 35
patient: 31
time: 27
doctor: 26
hospital: 25
wait: 24
service: 23
bad: 18
staff: 17
ask: 17

----- 10 most common 2-grams -----
speak badly: 8
service terrible: 5
doctor nurse: 4
long time: 4
emergency room: 3
feel bad: 3
doctor doctor: 3
poor service: 3
doctor staff: 3
nurse speak: 3

----- 10 most common 3-grams -----
ear nose throat: 2
patient official look: 2
official look happy: 2
look happy nurse: 2
doctor doctor wait: 2
doctor wait doctor: 2
wait doctor die: 2
poor service terrible: 2
service terrible thank: 2
terrible thank doctor: 2
```

Rama

word cloud for positive group

word cloud for negative group

Chula



Comment:
There are 4 negative words about x-ray.

Top 10 positive group

----- 10 most common 1-grams -----
good: 182
doctor: 176
hospital: 112
service: 103
patient: 82
nurse: 81
time: 64
care: 49
wait: 47
staff: 46

----- 10 most common 2-grams -----
good service: 34
doctor nurse: 31
service good: 25
doctor good: 22
long time: 18
wait long: 16
good care: 16
chula hospital: 16
use service: 13
nurse good: 13

----- 10 most common 3-grams -----
wait long time: 13
thank doctor nurse: 9
doctor nurse good: 8
chulalongkorn memorial hospital: 6
nurse good care: 6
king chulalongkorn memorial: 5
hospital good service: 5
good care patient: 5
pay attention patient: 5
service good doctor: 4

Top 10 negative group

----- 10 most common 1-grams -----
hospital: 40
come: 32
time: 29
wait: 29
nurse: 27
doctor: 26
good: 24
service: 23
bad: 21
patient: 19

----- 10 most common 2-grams -----
long time: 9
chula hospital: 6
waste time: 6
wait long: 5
sit wait: 5
chulalongkorn hospital: 4
doctor nurse: 4
x ray: 4
government hospital: 3
people want: 3

----- 10 most common 3-grams -----
wait long time: 5
patient wait long: 2
push car elevator: 2
want come service: 2
medical student good: 1
student good ask: 1
good ask symptom: 1
ask symptom examination: 1
symptom examination check: 1
examination check doctor: 1

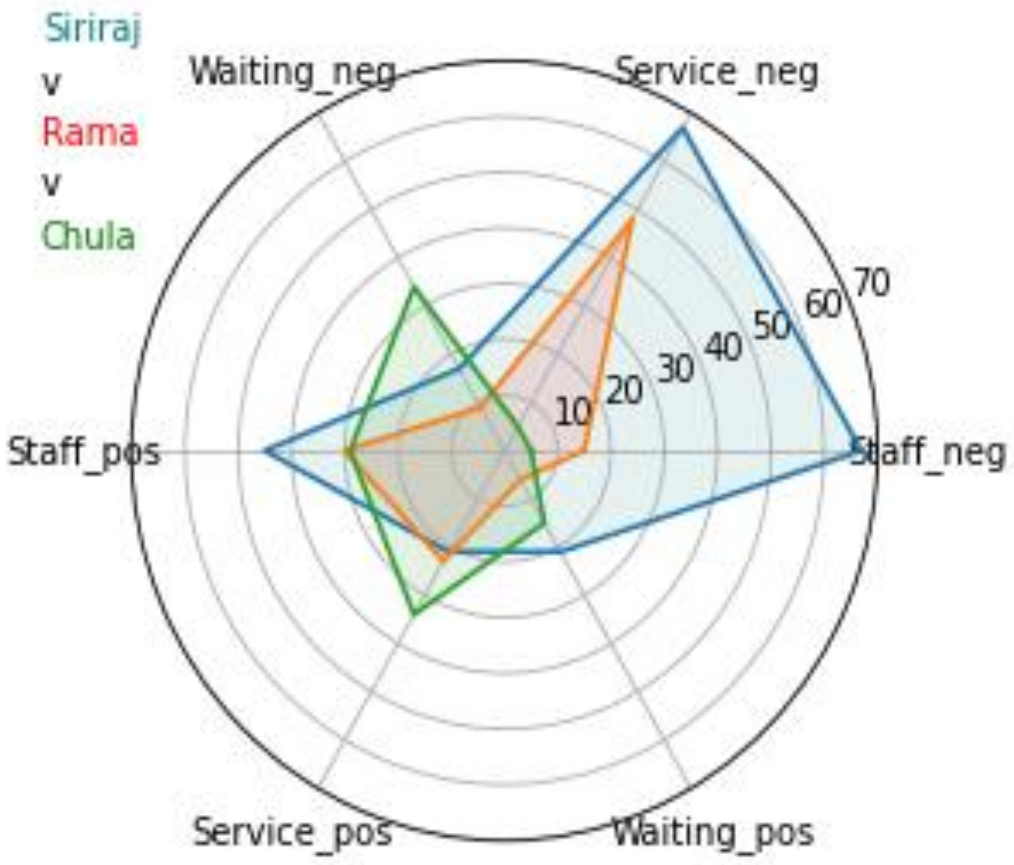
Chula

word cloud for positive group

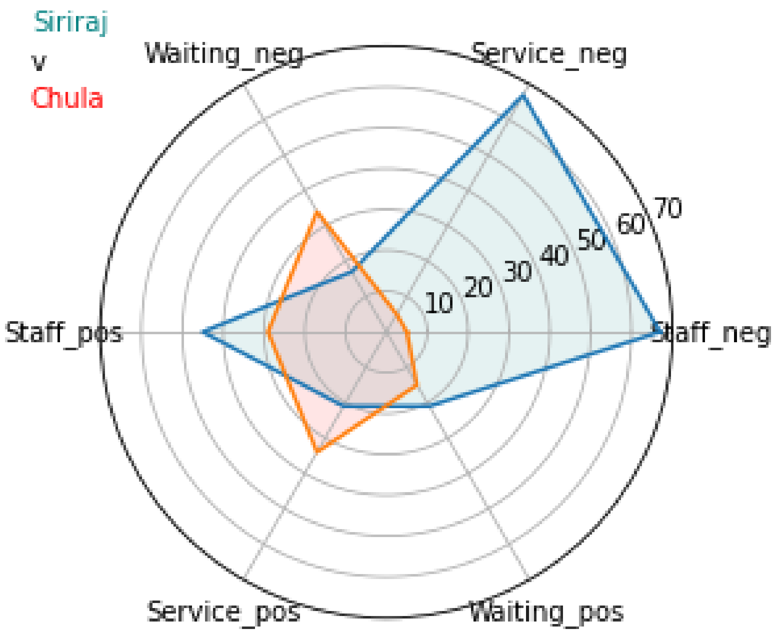
word cloud for negative group

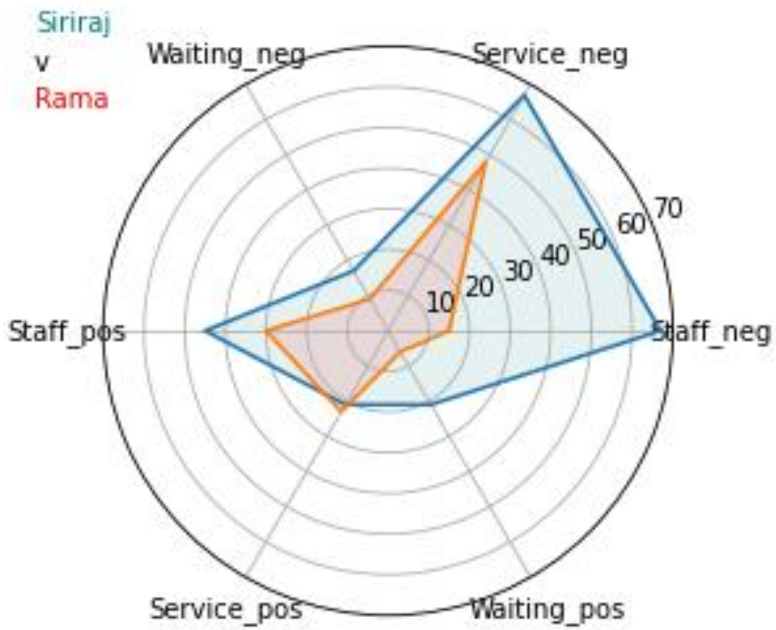
Comparisons: SI vs RA vs CU

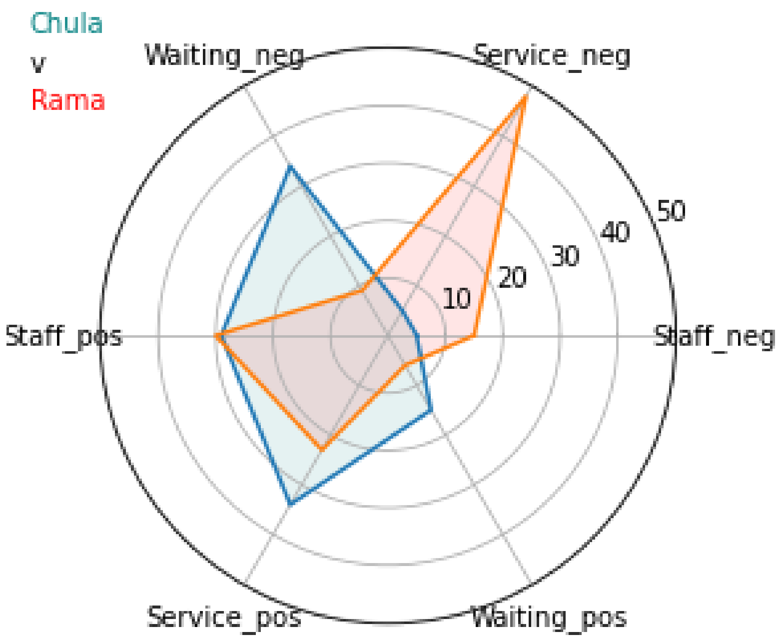
- Positive comments
  - All these 3 hospitals have quite similar ratios in each categories.
- Negative comments
  - Waiting: Chula (3x) > Siriraj (2x) > Rama (1x)
  - Service: Siriraj (7x) > Rama (5x)>> Chula (1x)
  - Staffs: Siriraj (7x)>> Rama (2x)> Chula (1x)
- Prescriptive analysis
  - Rama : should improve service quality
  - Siriraj: should improve service and staff
  - However, WE MUST LOOK IN DETAILS, that which particular topics in each dimension should be improved. FURTHER ANALYSIS IS NEEDED.



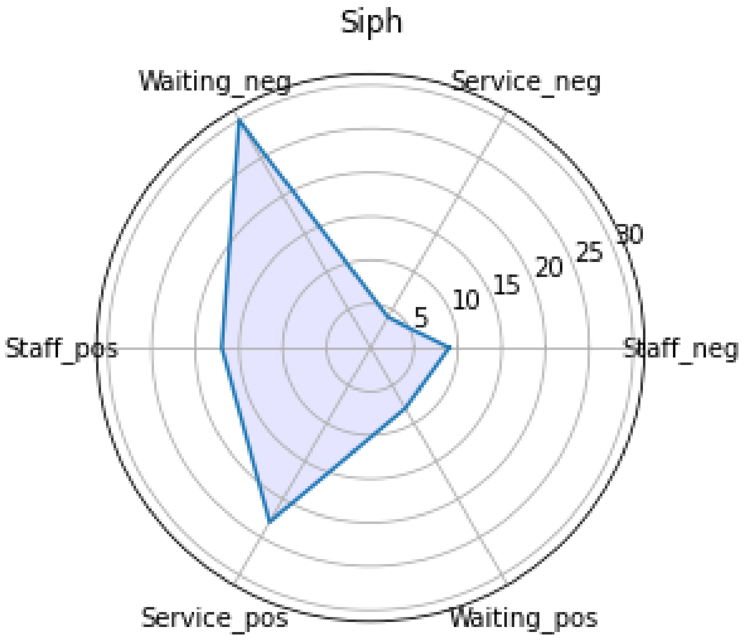
Pairwise comparisons







SiPH



Comment:
There are 39 positive words about clean.

Top 10 positive group

```
----- 10 most common 1-grams -----
good: 194
doctor: 139
service: 137
hospital: 103
nurse: 60
patient: 47
time: 45
siriraj: 44
treatment: 42
clean: 39

----- 10 most common 2-grams -----
good service: 40
service good: 23
doctor nurse: 20
use service: 14
good doctor: 14
doctor good: 13
long time: 12
good good: 10
wait long: 10
siriraj hospital: 10

----- 10 most common 3-grams -----
wait long time: 6
good service fast: 5
good service good: 5
good care hospital: 3
care hospital look: 3
hospital look clean: 3
look clean regular: 3
service doctor good: 3
siriraj piyamaharajkarun hospital: 3
hospital service good: 3
```

Top 10 negative group

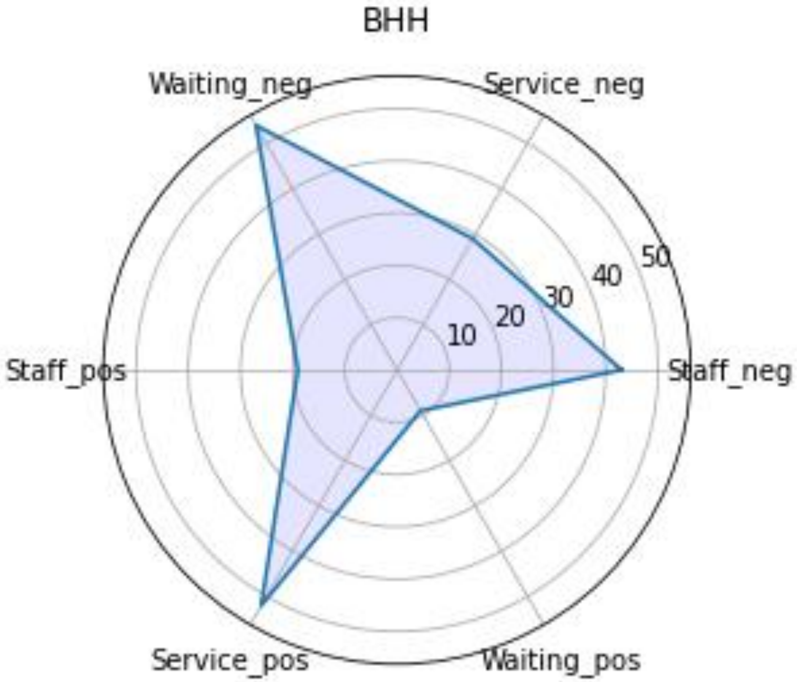
```
----- 10 most common 1-grams -----
doctor: 45
service: 36
wait: 36
patient: 30
nurse: 28
hospital: 27
good: 21
come: 20
time: 18
mother: 15

----- 10 most common 2-grams -----
long time: 5
doctor good: 4
wait queue: 4
wait hour: 4
wait doctor: 4
fever doctor: 3
service doctor: 3
doctor nurse: 3
patient wait: 3
time service: 3

----- 10 most common 3-grams -----
floor zone c: 2
mother fever doctor: 2
time service slow: 2
understand lot people: 2
room wait hour: 2
wait long time: 2
thai brother sister: 2
mother treat ischemic: 1
treat ischemic stroke: 1
ischemic stroke bedridden: 1
```

SiPH

word cloud for positive group

word cloud for negative group

BHH



Comment:
There are 9 negative words about expensive.

Top 10 positive group

----- 10 most common 1-grams -----
good: 158
service: 120
doctor: 105
hospital: 71
care: 45
patient: 38
nurse: 35
time: 33
like: 30
impressed: 28

----- 10 most common 2-grams -----
good service: 35
service good: 28
good care: 19
doctor good: 17
use service: 15
doctor nurse: 10
good hospital: 9
bumrungrad hospital: 9
service doctor: 8
hospital doctor: 8

----- 10 most common 3-grams -----
good service doctor: 6
wait long time: 5
good service attentive: 5
good hospital good: 4
good service fast: 4
hospital good service: 4
personally use service: 3
service good doctor: 3
use service feel: 3
thank good service: 3

Top 10 negative group

----- 10 most common 1-grams -----
time: 17
wait: 17
doctor: 12
long: 11
service: 10
expensive: 9
appointment: 7
check: 6
hour: 6
like: 6

----- 10 most common 2-grams -----
long time: 6
wait long: 4
long wait: 3
people use: 2
use service: 2
appointment time: 2
time person: 2
train manner: 2
check minute: 2
expensive mean: 2

----- 10 most common 3-grams -----
wait long time: 3
people use service: 2
doctor terrible price: 1
terrible price godly: 1
price godly service: 1
godly service generic: 1
service generic slow: 1
generic slow long: 1
slow long time: 1
long time lot: 1

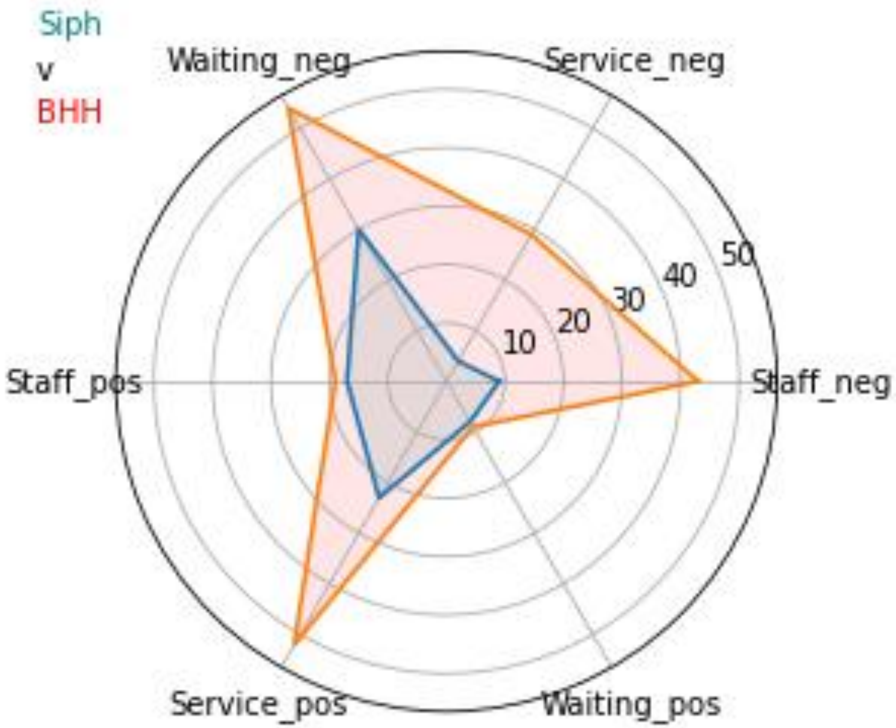
BHH

word cloud for positive group

word cloud for negative group

Comparisons: SiPH vs BHH

- SiPH performed better in every dimensions for negative comments.
- For positive comments, staff and waiting time are not quite different. But BHH shows far beyond SiPH in term of Services.
- Prescriptive Analysis
  - SiPH is doing well. But should be careful about waiting time that the peak is obvious, and it might be problems in the future. And should take care of service dimension for further competitive advantage.
  - BHH shows very extreme among every topics. Thus, it reflects the heterogeneity of the services, staff, and waiting time. BHH should explore in advance that why individuals have so much different experience e.g. patient’s expectation, etc. Anyway, BHH is still strong about service dimension among positive comments. If the negative received the same quality of the service as the positive did, the performance would be better.



Contents

DATA VISUALIZATIONT

- Web Scraping
- Original Data
- Data Preparation
- Data Visualization
- Choosing Hospital for Further analysis

TOPIC ANALYSIS

- NLP
- Bag of Words
- N-grams Model
- LDA Model
- Word Cloud
- Radar Chart

SENTINEL ANALYSIS

- Train and Test Dataset
- Model selection
- Hyperparameter tuning
- Model comparison

DISCUSSION

- Topic Analysis
- Sentinel Analysis
- Analysis of Error
- Limitation
- Suggestion

Train and Test Dataset

- Train : Test = 70:30

```
X_train.shape
(821,)
```

```
y_train.shape
(821,)
```

```
X_test.shape
(352,)
```

```
y_test.shape
(352,)
```

Model selection

- Logistic Regression
  - It is a basic model (still used in classical modelling and machine learning) that I want to use as baseline model.
  - The predicted outcomes are 0 and 1 (positive and negative sentinel)
- SVC (Support Vector Classification)
  - Similar to support vector machine
  - Better for small dataset but a bit complex
- MNB (Multinomial Naive Bayes classifier)
  - Probabilistic features suit for NLP techniques
  - Possible for more than two outcomes
    (not in this model but possible in real analysis)
  - Assumption of independence among features

Logistic Regression

Pros

- Simple to understand and explain
- It seldom overfits
- Using L1 & L2 regularization is effective in feature selection
- The best algorithm for predicting probabilities of an event
- Fast to train
- Easy to train on big data thanks to its stochastic version

Cons

- You have to work hard to make it fit nonlinear functions
- Can suffer from outliers

SVC (Support Vector Classification)

Pros

- Automatic nonlinear feature creation
- Can approximate complex nonlinear functions

Cons

- Difficult to interpret when applying nonlinear kernels
- Suffers from too many examples, after 10,000 examples it starts taking too long to train

MNB (Multinomial Naive Bayes classifier)

Pros

- Computationally fast
- Simple to implement
- Works well with small datasets
- Works well with high dimensions
- Perform well even if the *Naive Assumption* is not perfectly met. In many cases, the approximation is enough to build a good classifier.

Cons

- Require to remove correlated features because they are voted twice in the model and it can lead to over inflating importance.
- If a categorical variable has a category in test data set which was not observed in training data set, then the model will assign a zero probability. It will not be able to make a prediction.

Logistic Regression

- Model with the best hyperparameter tuning with Grid Search

```
#Model with the best parameters
clf_lgr_best = Pipeline([
      ('tfidf', TfidfVectorizer(ngram_range=(1,2),use_idf=False)),
      ('clf', LogisticRegression(C=10 , penalty='l2' ,verbose=1,n_jobs=-1))])
```

- Model performance

```
              precision    recall  f1-score   support

         0.0       0.74      0.42      0.54        40
         1.0       0.89      0.97      0.93       195

    accuracy                           0.88       235
   macro avg       0.82      0.70      0.73       235
weighted avg       0.87      0.88      0.86       235

[[ 17  23]
 [  6 189]]
```

SVC (Support Vector Classification)

- Model with the best hyperparameter tuning with Grid Search

```
#Model with the best parameters
clf_svc_best = Pipeline([
      ('tfidf', TfidfVectorizer(ngram_range=(1,1),use_idf=True)),
      ('clf', SVC(C=10, verbose=1))])
```

- Model performance

```
              precision    recall  f1-score   support

         0.0       0.00      0.00      0.00        40
         1.0       0.83      1.00      0.91       195

    accuracy                           0.83       235
   macro avg       0.41      0.50      0.45       235
weighted avg       0.69      0.83      0.75       235

[[  0  40]
 [  0 195]]
```

MNB (Multinomial Naive Bayes classifier)

- Model with the best hyperparameter tuning with Grid Search

```
#Model with the best parameters
clf_mnb_best = Pipeline([
      ('tfidf', TfidfVectorizer(ngram_range=(1,2),use_idf=False)),
      ('mnb', MultinomialNB( alpha=0.5, fit_prior=False, class_prior=None))])
```

- Model performance

```
              precision    recall  f1-score   support

         0.0       0.57      0.10      0.17        40
         1.0       0.84      0.98      0.91       195

    accuracy                           0.83       235
   macro avg       0.71      0.54      0.54       235
weighted avg       0.80      0.83      0.78       235

[[  4  36]
 [  3 192]]
```

The best Model

- The best model from this dataset is Logistic Regression with Accuracy of 0.88
- However, MNB model also performed quite well (accuracy 0.83 compared to 0.88 of logistic regression model). If we have more data, in my opinion MNB should be the best model because it works well with large dimensional data, fast computation, and in real word NLP that may classify more than 2 outcomes. Even though, there is ‘Zero Frequency’ problem but it can be solved with smoothing technique such as Laplace smoothing, and it works well.

Contents

DATA VISUALIZATIONT

- Web Scraping
- Original Data
- Data Preparation
- Data Visualization
- Choosing Hospital for Further analysis

TOPIC ANALYSIS

- NLP
- Bag of Words
- N-grams Model
- LDA Model
- Word Cloud
- Radar Chart

SENTINEL ANALYSIS

- Train and Test Dataset
- Model selection
- Hyperparameter tuning
- Model comparison

DISCUSSION

- Topic Analysis
- Sentinel Analysis
- Analysis of Error
- Limitation
- Suggestion

Error analysis

- Misclassifications of the model (both logistic and MNB are fully shown in excel files) are mainly due to
  - Too short comments

Index	Comment	Score	Hospital	**Translator**	Sentiment	Post-NLP	Prediction
53	สะอาด บริการดีค่ะ	4	โรงพยาบาลศิริราช	Clean, good service	1	clean good service	0
122	พ่อผมไม่สบาย	1	โรงพยาบาลรามาธิบดี	My father is sick.	0	father sick	1

  - Unable to translate from google translator
    e.g. มากกกกๆ นานเกิ๊น
  - Sarcastic word
    e.g. ก็ดี (ไม่ได้หมายความว่าดีจริงๆ) ดีมาก!!

Error analysis

- Misclassifications of the model (both logistic and MNB are fully shown in excel files) are mainly due to
  - Real misclassification

Index	Comment	Score	Hospital	**Translator**	Sentiment	Post-NLP	Prediction
80	พยาบาลเอาใจใส่ดี พูดจาสุภาพ คุณหมอก็แนะนำการดูแลตัวเองอย่างละเอียด	5	โรงพยาบาลศิริราช	The nurse was very attentive, polite, and the doctor advised him to take good care of himself.	1	nurse attentive polite doctor advise good care	0
130	บริการแย่ โดยเฉพาะพวกพยาบาล	1	โรงพยาบาลรามาธิบดี	Poor service, especially for nurses	0	poor service especially nurse	1

  - No number after NLP process
  - Losing of negative word after NLP process

Limitation

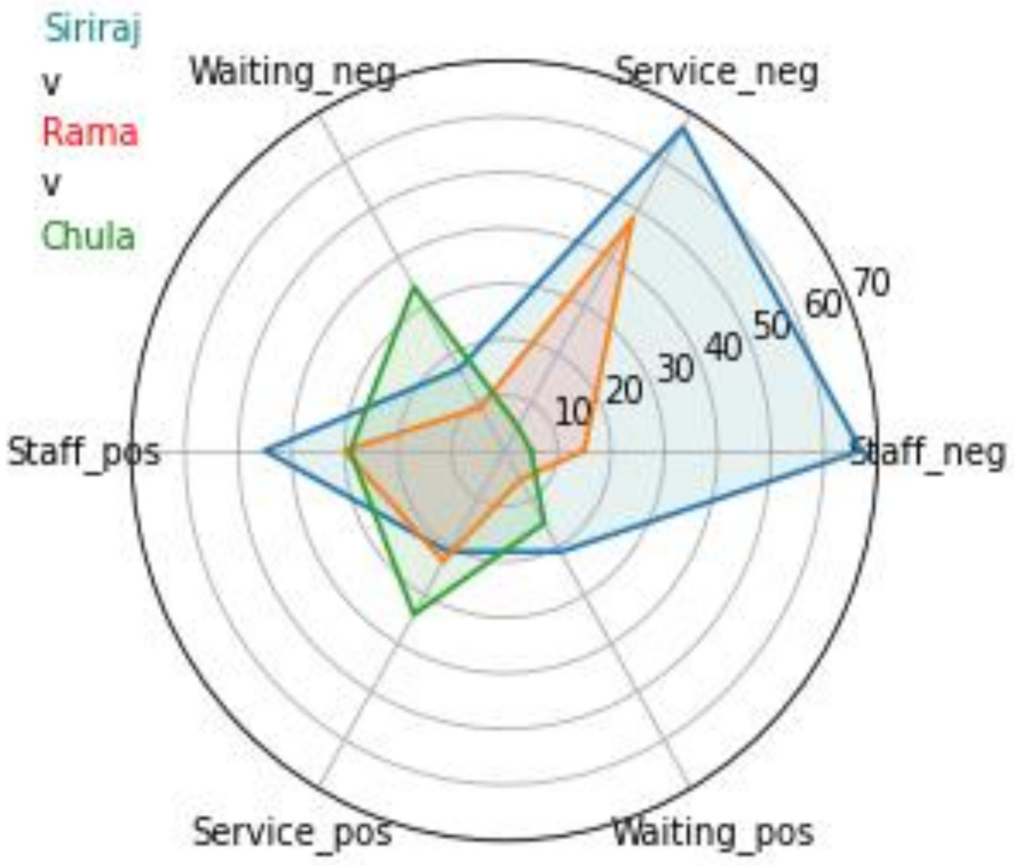
- Both good and bad topics in ONE comment
- Small number of data
- English translator from Thai
  - Mistranslation
- Special cases
  - Sarcastic word
  - Slang
- Low quality of hyperparameter tuning process
  - Hardware

Suggestion

- Use larger sample size
- Try to use Thai NLP to decrease translation problem but might deal with poorer performance.
- For topic analysis, should look further in details for example
  - People who complain about waiting time are the ones who did not use mobile apps.
  - Characteristics of Doctors and Nurses in Both positive and negative groups. E.g. Speak badly, being so late, good care, clear instruction what to do, etc. These will help a lot.
- Better technique of modeling and hyperparameter tuning including deep learning such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) with/without Transfer learning.

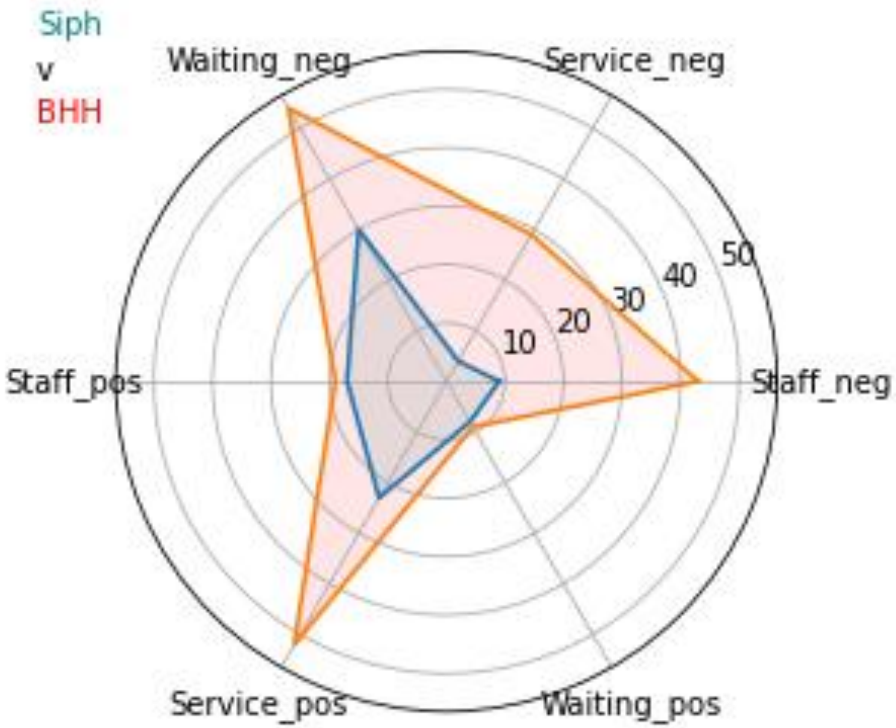
Comparisons: SI vs RA vs CU

- Positive comments
  - All these 3 hospitals have quite similar ratios in each categories.
- Negative comments
  - Waiting: Chula (3x) > Siriraj (2x) > Rama (1x)
  - Service: Siriraj (7x) > Rama (5x)>> Chula (1x)
  - Staffs: Siriraj (7x)>> Rama (2x)> Chula (1x)
- Prescriptive analysis
  - Rama : should improve service quality
  - Siriraj: should improve service and staff
  - However, WE MUST LOOK IN DETAILS, that which particular topics in each dimension should be improved. FURTHER ANALYSIS IS NEEDED.



Comparisons: SiPH vs BHH

- SiPH performed better in every dimensions for negative comments.
- For positive comments, staff and waiting time are not quite different. But BHH shows far beyond SiPH in term of Services.
- Prescriptive Analysis
  - SiPH is doing well. But should be careful about waiting time that the peak is obvious, and it might be problems in the future. And should take care of service dimension for further competitive advantage.
  - BHH shows very extreme among every topics. Thus, it reflects the heterogeneity of the services, staff, and waiting time. BHH should explore in advance that why individuals have so much different experience e.g. patient's expectation, etc. Anyway, BHH is still strong about service dimension among positive comments. If the negative received the same quality of the service as the positive did, the performance would be better.



THANK YOU

24Slides