Sentiment Analysis Over Time

Machine Learning Complements

Group C

Bruna Marques, up202007191
Diogo Silva, up202004288
Lia Vieira, up202005042
Pedro Fonseca, up202008307

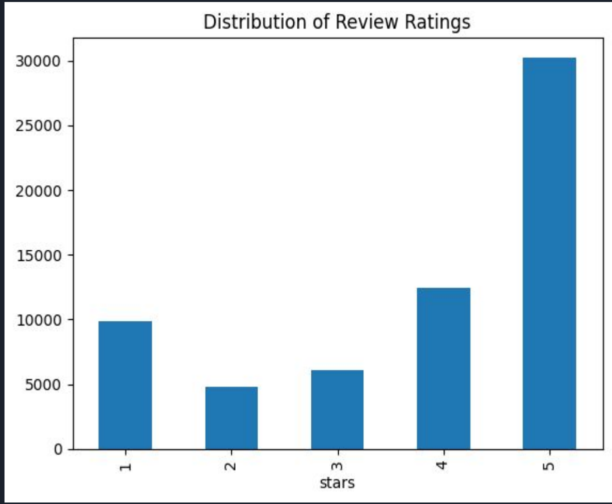
Domain Description

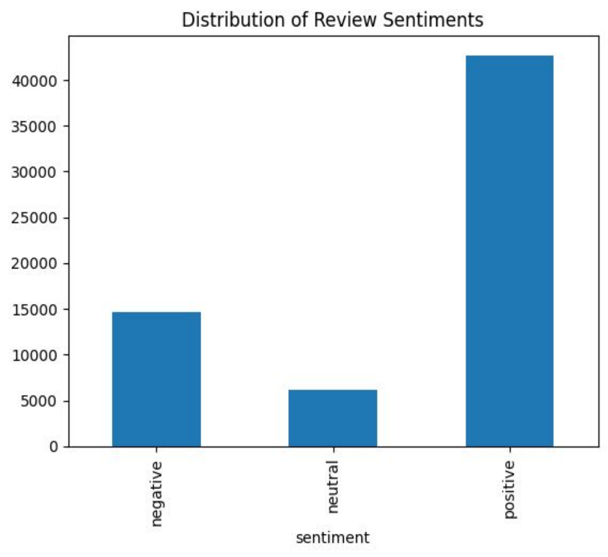
- **Yelp dataset** - businesses, check-ins, reviews, tips, users
- Implement and evaluate **NLP** for **sentiment analysis** to predict whether reviews express positive, negative, or neutral sentiments
- Track its evolution **over time**, identifying relevant **patterns**

Natural Language Processing

Exploratory Data Analysis

- Reviews categorized based on their **star** ratings:
    - 1 or 2 stars - **negative**
    - 3 stars - **neutral**
    - 4 or 5 stars - **positive**





- Positive ratings: almost **70%** of all data
- Sentiment value: 1 for positive reviews, 0 for neutral and -1 for negative

Data Preparation

- Processed each review text:
  a. Removed non-alphabetic characters using a regular expression:
     - except from '!' - considered essential for expressing strong emotions
  b. Converted to lowercase
  c. Performed tokenization
  d. Prepended **NOT_** to words affected by negation tokens – n't, not, no, never, etc
  e. Removed stopwords
  f. Performed lemmatization

Feature Engineering

- **Vectorized representations of words:**
  - sparse - BoW, 1-hot encoding, N-grams, TF-IDF
  - dense - word2vec

- **Dealing with negation**: Lexicons
  - 2 new columns, **positive** and **negative**:
    - record count of positive and negative words, respectively
  - positive and negative word sets from NLTK to identify and count these words
  - combine this approach with the bag-of-words to enhance feature representation

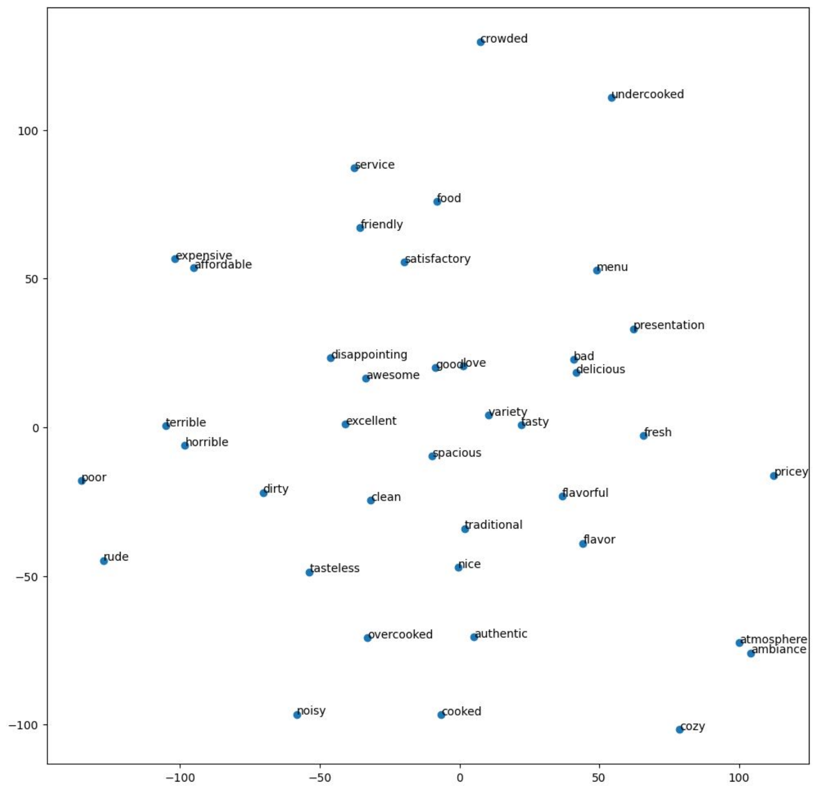
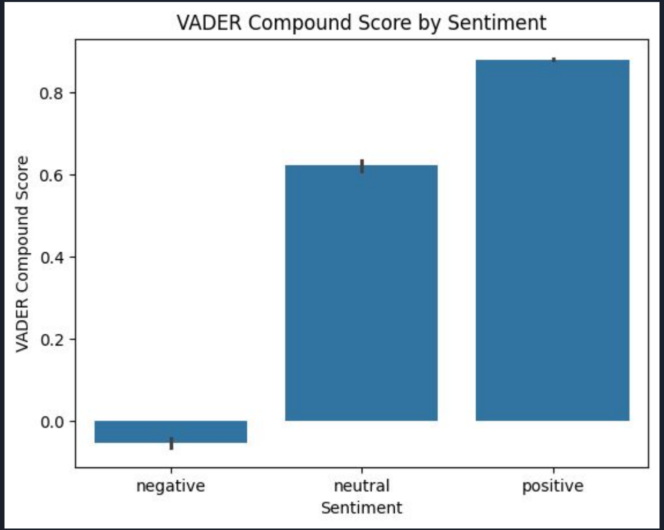


Fig. Word2vec vector visualization

Modelling - VADER

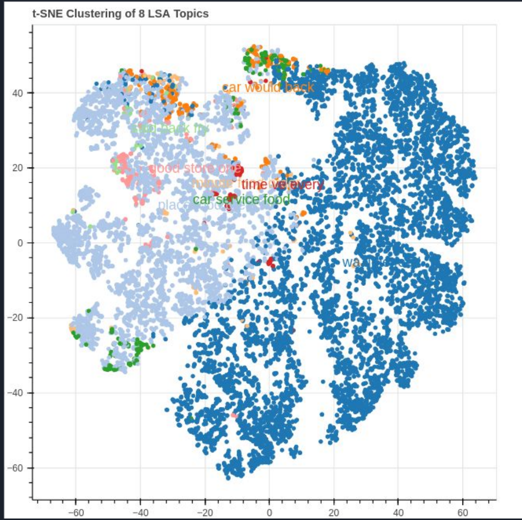
- Emotional tone evaluation
- Individual scores for positive, neutral, and negative sentiment
- Normalized, weighted composite score
- Compound score rounded to -1, 0, or 1 to indicate **negative**, **neutral**, or **positive** sentiment
- Broader range for neutral sentiment (-0.5 to 0.5):
    - higher number of reviews categorized as **neutral**



Modelling - Unsupervised

Topic Modelling

- **Latent Semantic Analysis (LSA)**
    - Identified 8 topics within the dataset and how frequently they appear.
    - For each topic, extracted and the **top 10** most common words.
    - Topics revolve around food quality and services, customer service and communication, and hotel accommodations.



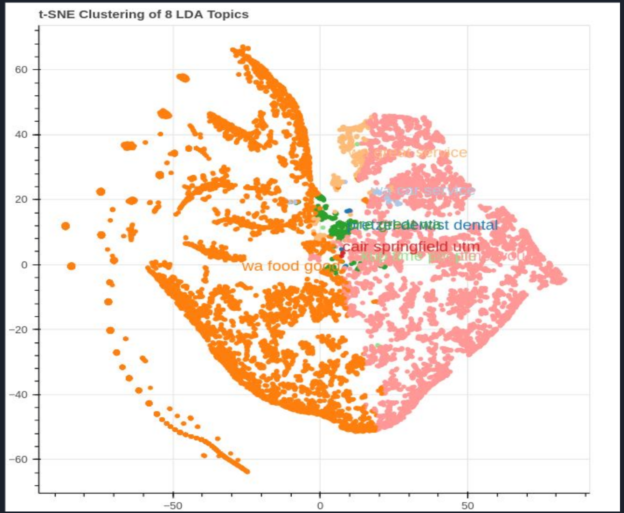
Topic 1: food place good great time order service get like go
Topic 2: call would told car room day said get time back
Topic 3: order wait food minut time ask servic get got call
Topic 4: cream ice flavor chicken like good cheese sauce tasty

...

Modelling - Unsupervised

Topic Modelling

- **Latent Dirichilet Allocation (LDA)**
  - Identified 8 topics within the dataset and how frequently they appear.
  - For each topic, extracted and the **top 10** most common words.
  - Topics revolve around preferences, food quality, prices, and overall satisfaction



Topic 1: store locat one go get shop time item like always
Topic 2: room hotel stay one would clean staff night get bed
Topic 3: good order food chicken fry sauce great like cheese
Topic 4: food place good like great time get go one order

...

Modelling

- We considered the stars ratings in reviews as our ***Ground Truth***.
  - **Limitation**: Users sometimes might write highly positive reviews but still assign a lower rating and vice-versa.
- Splitted the data into training and test sets (80-20)
- Applied **Cross Validation** on all classifiers using the Word2vec matrix
  - None of the models exhibit signs of severe overfitting based on the cross-validation results
  - SVM was the top performer in terms of accuracy, but all models exhibited moderate precision, recall, and F1 scores
- Hyperparameter tuning through **GridSearchCV** using the Word2Vec matrix
  - Crucial for improving the classifiers' performance.

	Decision tree	**Random forest**	**SVM**	**Logistic Regression**	**Perceptron**	**XGB**
Best F1-score	0.53	0.62	0.70	0.72	0.70	0.69
Best parameters	max-depth:10	max-depth: None n-estimators: 200	C: 1.0 kernel: rbf	C: 0.1 penalty: l2 solver: liblinear	alpha: 0.0001 penalty: None	learning-rate: 0.1 max-depth: 5 n-estimators: 200

Modelling - Supervised

- Balanced the training set using **RandomOverSampler** for oversampling
  - Leading to more reliable performance metrics
- Each classifier trained using the **best hyperparameters** obtained from the grid search
- Models evaluated using accuracy, precision, recall, and F1 score.
- Evaluation of multiple classifiers for sentiment analysis:

MultinomialNB

	N-grams	**Tf-idf**
Precision	0.69	0.54
Recall	0.71	0.59
F1	0.70	0.56
Accuracy	0.78	0.79

Decision Tree

	BoW with negation	**Word2vec**
Precision	0.61	0.43
Recall	0.61	0.42
F1	0.58	0.41
Accuracy	0.68	0.56

Modelling - Supervised

Random Forest

	BoW with negation	Word2vec
Precision	0.79	0.67
Recall	0.61	0.47
F1	0.58	0.46
Accuracy	0.81	0.66

SVM

	Tf-idf	Word2vec
Precision	0.80	0.84
Recall	0.64	0.59
F1	0.62	0.54
Accuracy	0.83	0.76

Logistic Regression

	One hot	Word2vec
Precision	0.72	0.60
Recall	0.68	0.59
F1	0.69	0.59
Accuracy	0.82	0.72

Modelling - Supervised

Perceptron

	N-grams	Word2vec
Precision	0.70	0.58
Recall	0.68	0.57
F1	0.69	0.57
Accuracy	0.81	0.70

XGB

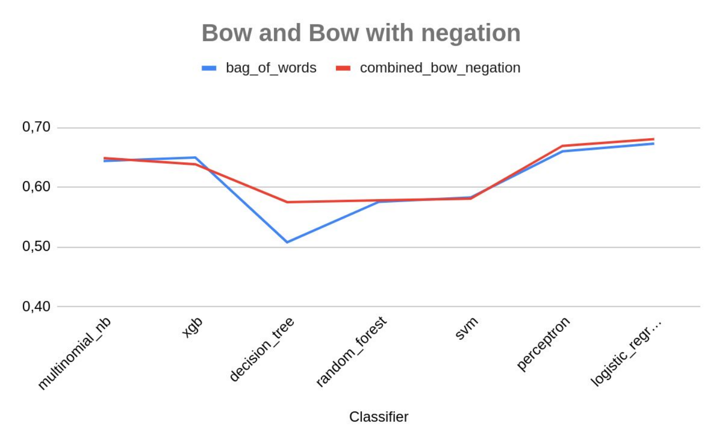
	Tf-idf	Word2vec
Precision	0.74	0.61
Recall	0.65	0.53
F1	0.65	0.51
Accuracy	0.81	0.72

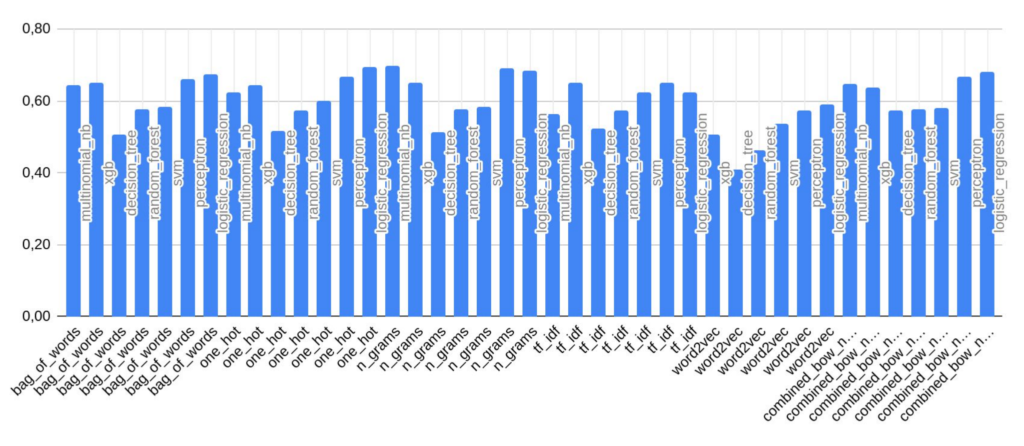
Conclusions

- **Topic modelling** is not a good strategy in this case since the data is very similar, so the resulting topics are not well distributed.
- Due to the significant **imbalance** in the dataset, undersampling would lose valuable information and **oversampling** would result in many duplicated data.
- **F1 score** was the primary metric used to compare and evaluate the performance of the classifiers.
- The performance of classifiers varies significantly depending on the feature representation used.

Conclusions II

- Word2Vec embeddings generally performs worse, particularly with the **Decision Tree** classifier.
- Combination of BoW with *handling negation* slightly improved the performance of Bag of words.
- Overall, the algorithms that showed better performance were **Logistic Regression** and **Perceptron**.



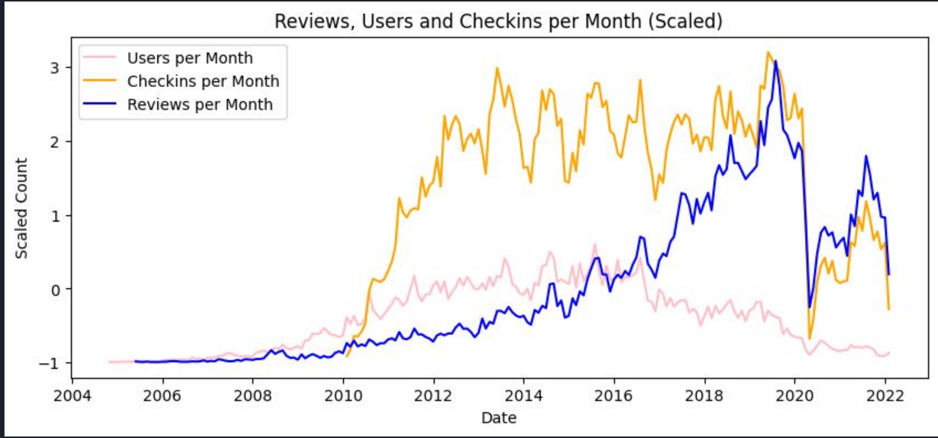


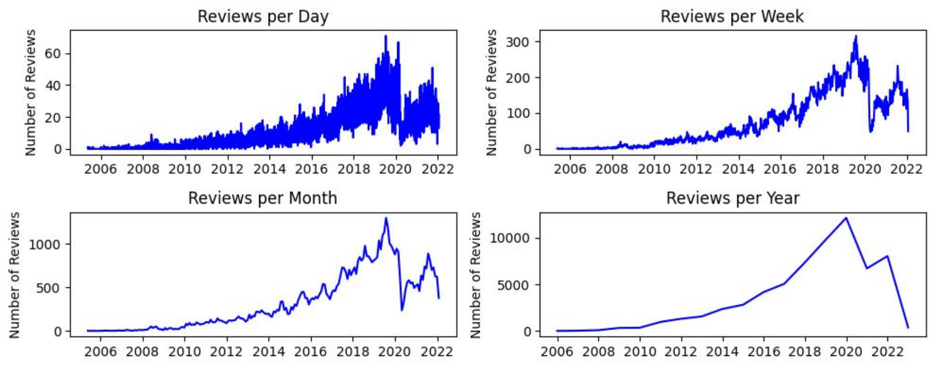
Time-Series Analysis

Data Preparation

Exploration

- Different periods visualized (daily, weekly, monthly, yearly):
    - **monthly** more suitable
- Data from dataset with **dates** - users, checkins, reviews:
    - increase in users and checkins - increase in reviews
    - **reviews before 2010 discarded** - not enough users for significant sentiment

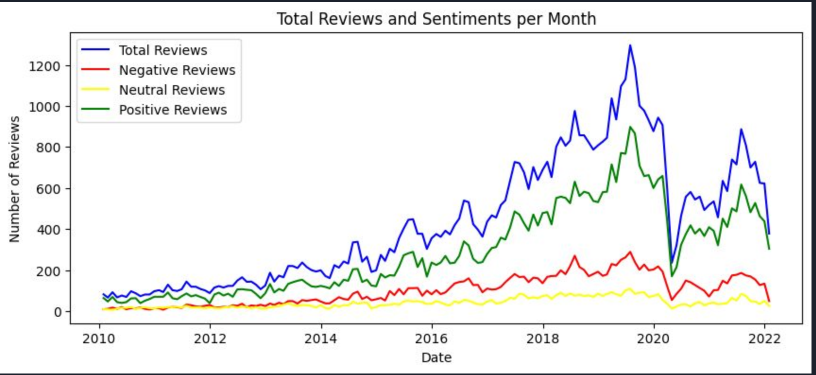


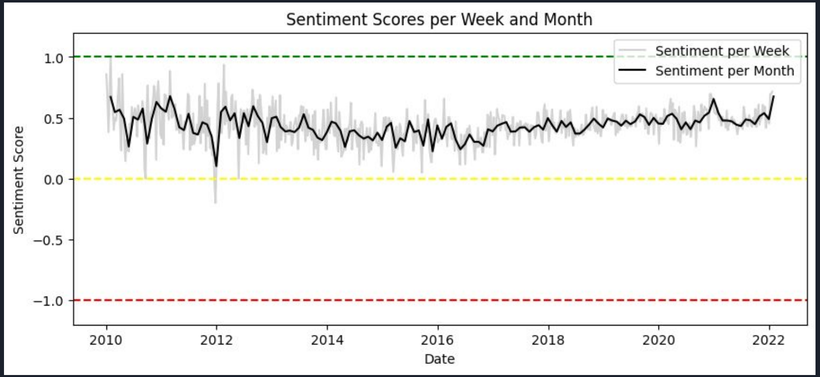


Data Preparation

Exploration

- Review sentiment approaches:
  - sentiment values (-1, 0, 1) based of review stars
  - **3 time series** - 1 for each sentiment:
    - visible trend, unclear seasonality/cycle
    - outliers - COVID
    - varying mean/variance
  - single time series - **monthly sentiment mean**:
    - unlikely trend, unlikely seasonality/cycle
    - some variance but not extreme
  - monthly sentiment mean time series for **3 major business categories (Restaurants, Nightlife, B&B)**
  - dealing with **missing values** - rolling mean (3)

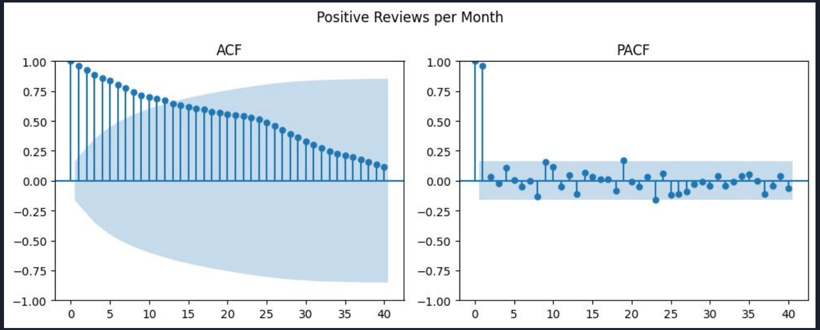


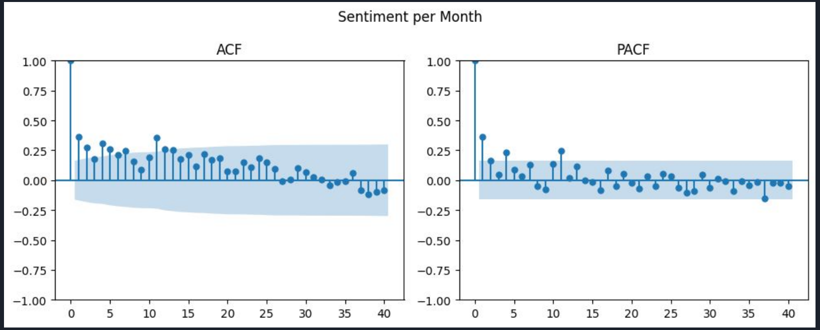


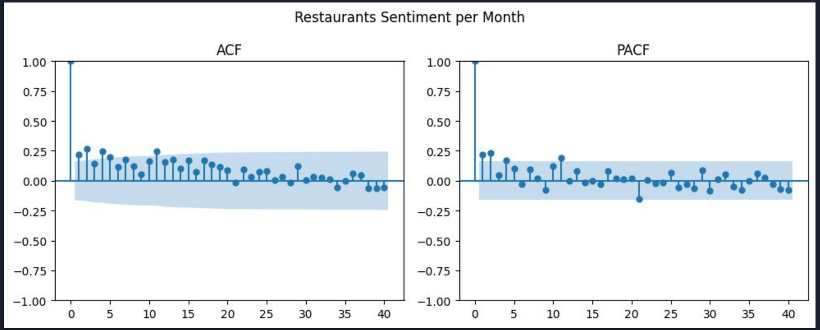
Data Preparation

ACF & PACF

- Autocorrelation in **negative, neutral, positive reviews**:
  - visible trend - large and positive in small lags
  - absence of seasonality or cyclic pattern
- Autocorrelation in **monthly sentiment mean**:
  - no strong trend
  - no apparent seasonality or cyclic pattern
- Autocorrelation in **restaurant sentiment**:
  - no strong trend
  - no seasonality or cyclic pattern



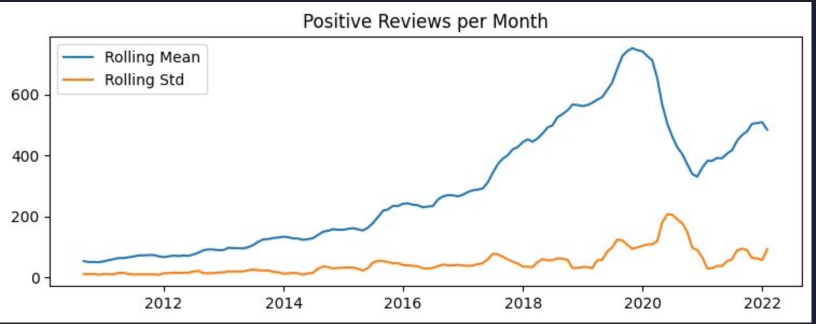


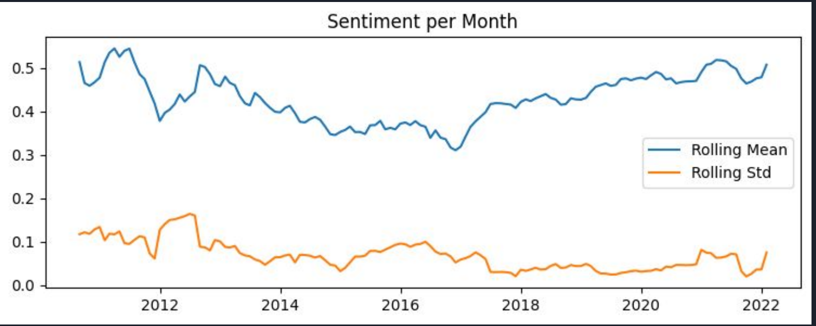


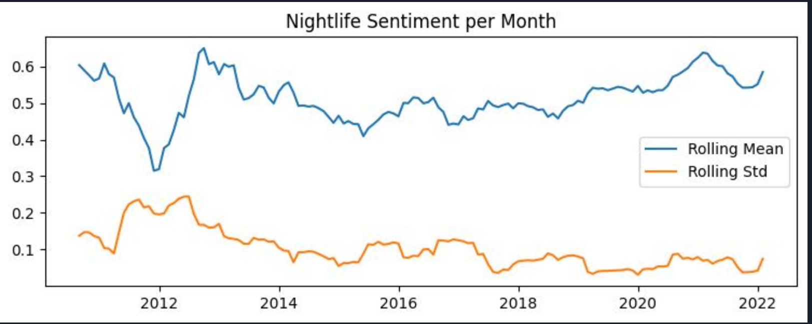
Data Preparation

Stationarity

- ADF, KPSS and PP tests
- Visualization - rolling mean and standard deviation
- Stationarity in **negative, neutral, positive reviews**:
  - tests suggest no stationarity
  - varying mean - non-stationary
- Stationarity in **monthly sentiment mean**:
  - inconsistent test results
  - varying mean (not much) - maybe non-stationary
- Stationarity in **nightlife sentiment**:
  - tests suggest stationarity
  - likely stationary

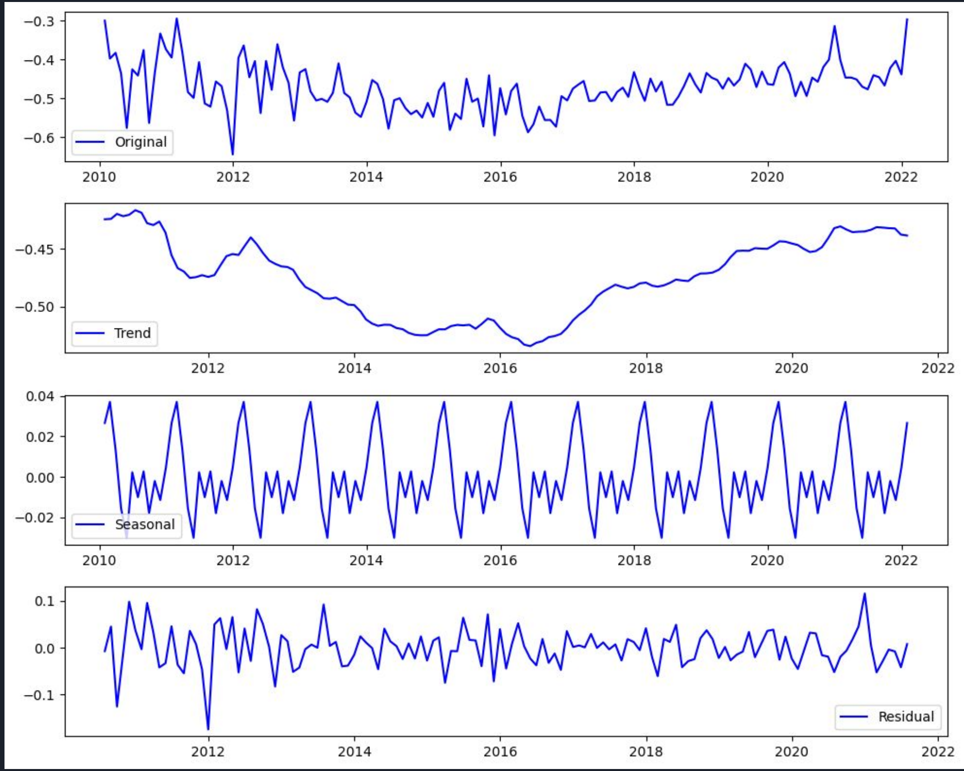


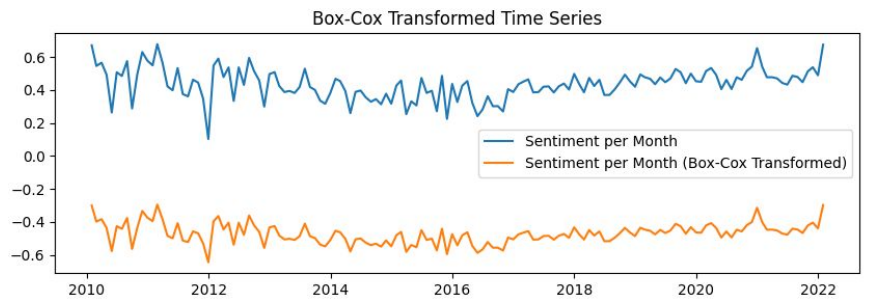




Feature Engineering

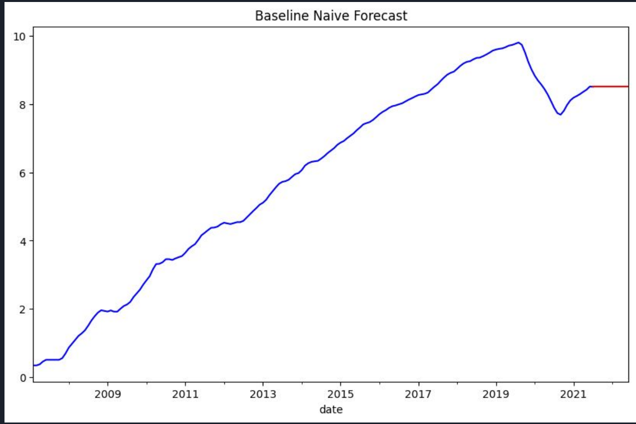
- Trends are roughly linear:
  - **box-cox transformation with summed constant** (stabilize variance) + **additive decomposition**





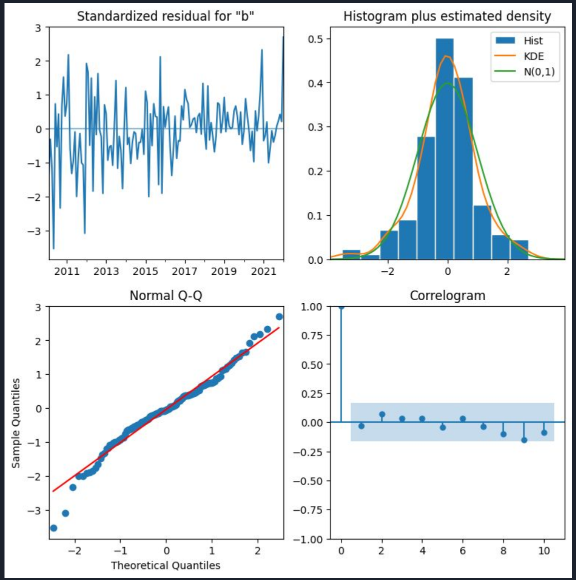
Modelling

- **Cross-Validation** to evaluate our models
- Minimum training set - **24 points** (2 years)
- We tried modelling with: **Original Sentiment** & **Box-Cox**
  - reverting forecast from models with box-cox to original data
- **Metrics** used: **MAE**, **MSE**, **RMSE**, **MAPE**, **MASE**
- **MSE** (mean squared error) - to decide best models
- **Baseline** models:
  - useful for establishing a performance baseline against complex models
  - mean, median and naive strategies used

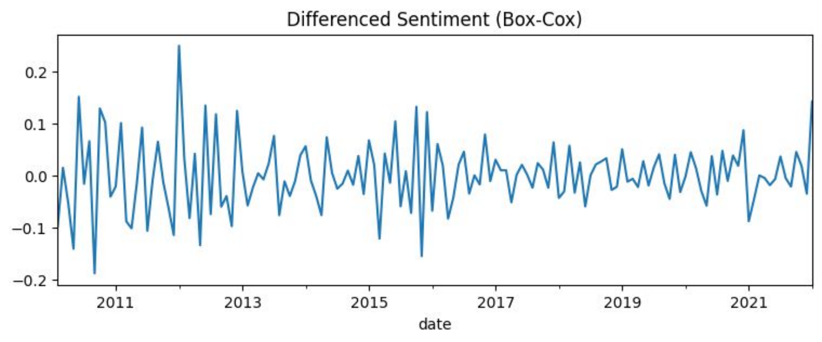


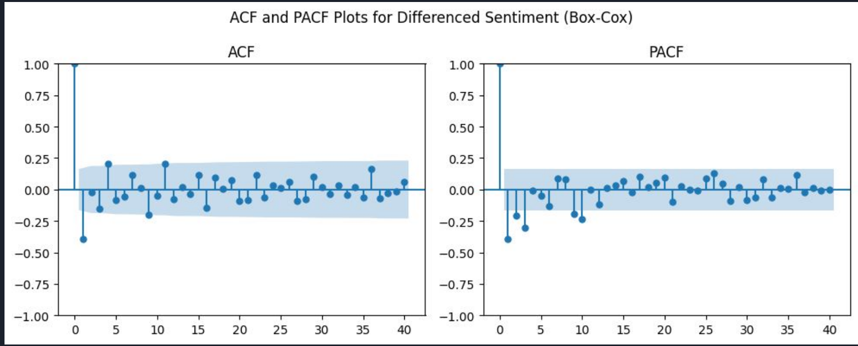
Modelling

ARIMA



- **Monthly sentiment mean** and **sentiment per category**
- Data from box-cox (stabilized variance):
  - differencing - stabilize mean
  - verify **stationarity** - ADF test, data visualization
- **ACF/PACF** interpretation
- Model fitting with **AIC** - order values
- **Residuals** diagnostics





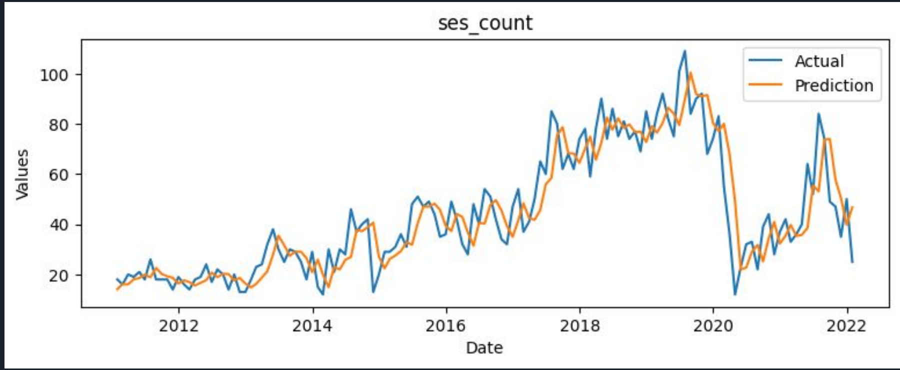
Modelling

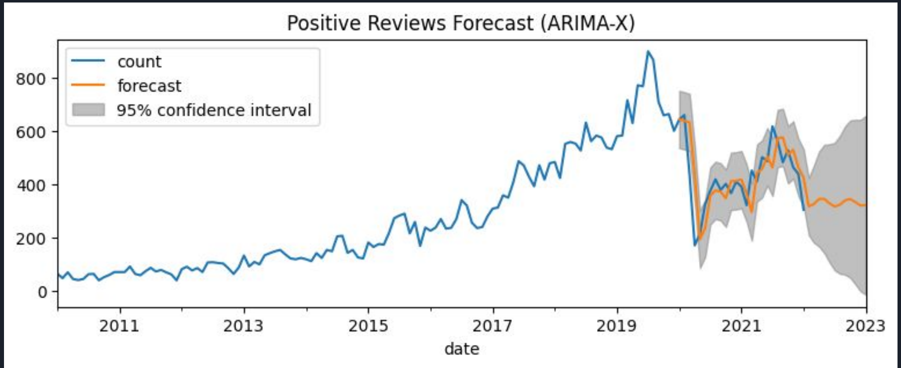
- **ARIMA-X**:
  - negative, neutral and positive review - monthly counts
  - exogenous variable for COVID dates (1st half of 2020):
    - significant event with lasting impact on data
    - definition of variable preferable to preprocessing
- **Simple Exponential Smoothing**:
  - useful for series with no clear trend or seasonal pattern:
    - mean sentiment, mean sentiment per category
- **Holt's Linear (Damped) Trend**:
  - adequate for forecasting of data with a trend:
    - negative/neutral/positive review counts

Model Evaluation

Negative/Neutral/Positive Review Counts

- **Negative reviews**:
  - best model - **Baseline Naive**
  - possibly due to COVID:
    - sudden fall made all other models incapable of following the trend
- **Neutral reviews**:
  - best model - **Simple Exponential Smoothing**
  - weak trend - lack of reviews
- **Positive reviews**:
  - best model - **Baseline Naive**
  - exponential smoothing failed to perform

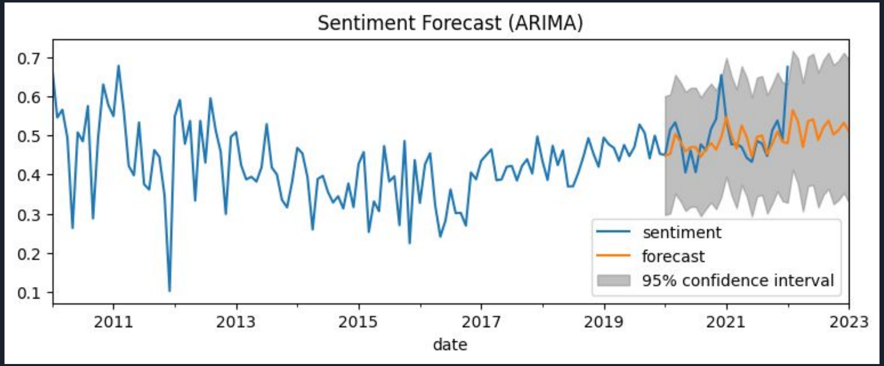


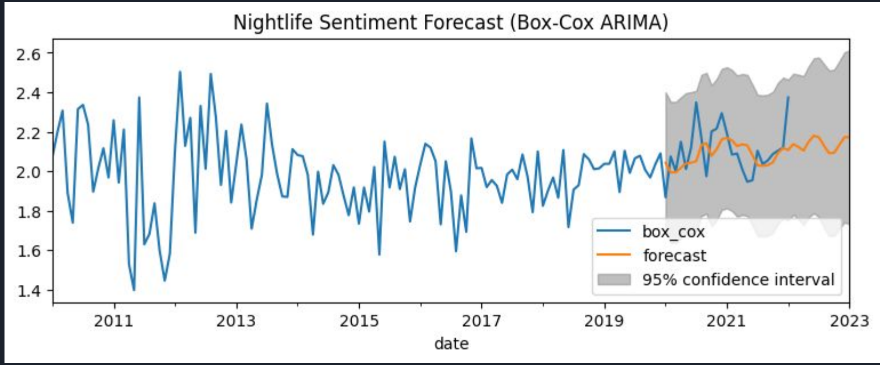


Model Evaluation

Total/Restaurant/Nightlife/B&B Mean Sentiment (monthly)

- **Total mean sentiment**:
  - best model - **ARIMA**
  - good order selection/preprocessing
- **Restaurant mean sentiment**:
  - best model - **Simple Exponential Smoothing**
  - no clear trend - not a lot of reviews
- **Nightlife mean sentiment**:
  - best model - **baseline mean**
  - some stationarity in original time series
- **B&B mean sentiment**:
  - best model - **ARIMA**



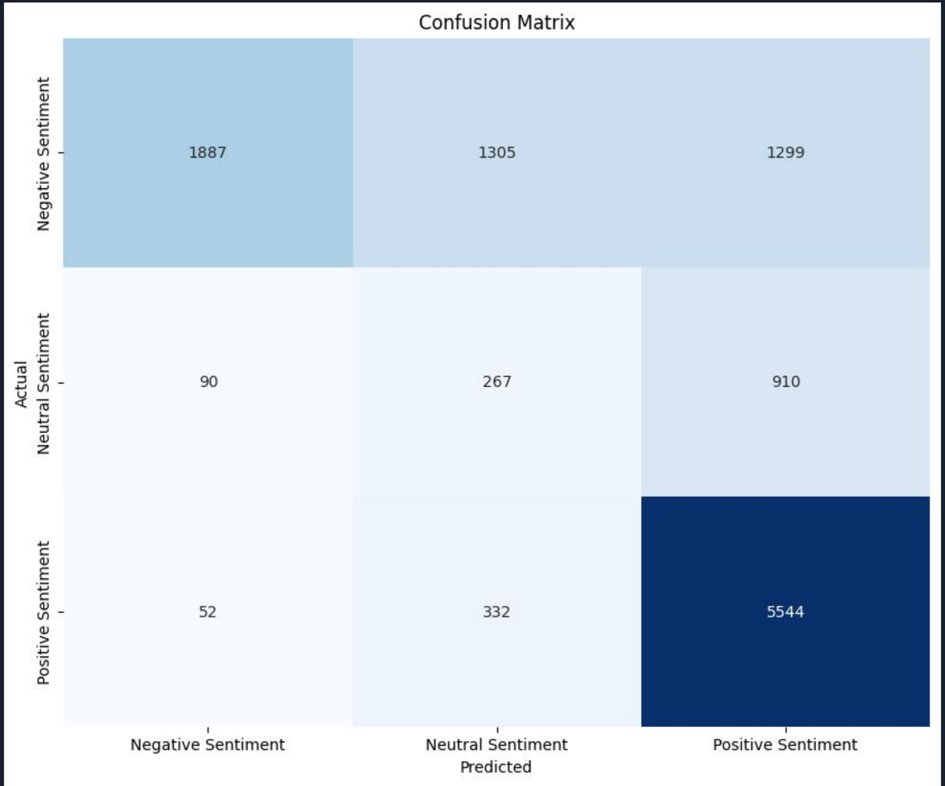


Conclusions

- Forecasting for **negative/neutral/positive review counts**:
  - difficult challenge - high chance of error
  - even ARIMA-X (with COVID variable) did not fulfill expectations:
    - possibly due to major COVID impact on data
- ARIMA models:
  - using **box-cox** had worst metric results **vs** using original series:
    - struggles in finding the perfect parameters
- **Future work**:
  a. use additive decomposition components - trend, seasonal, residuals
  b. model them individually
  c. forecast for original dataset summing the trend/seasonal/residual predictions

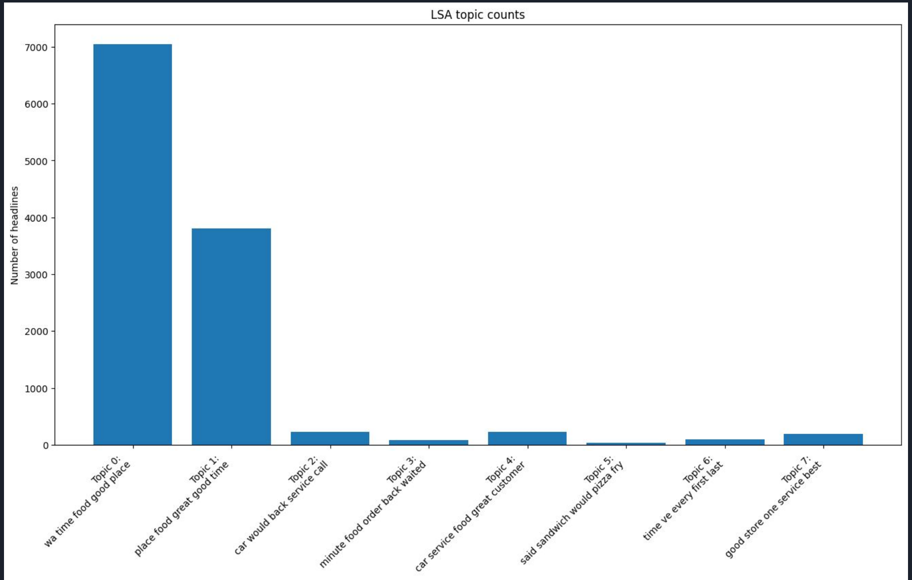
ANNEXES

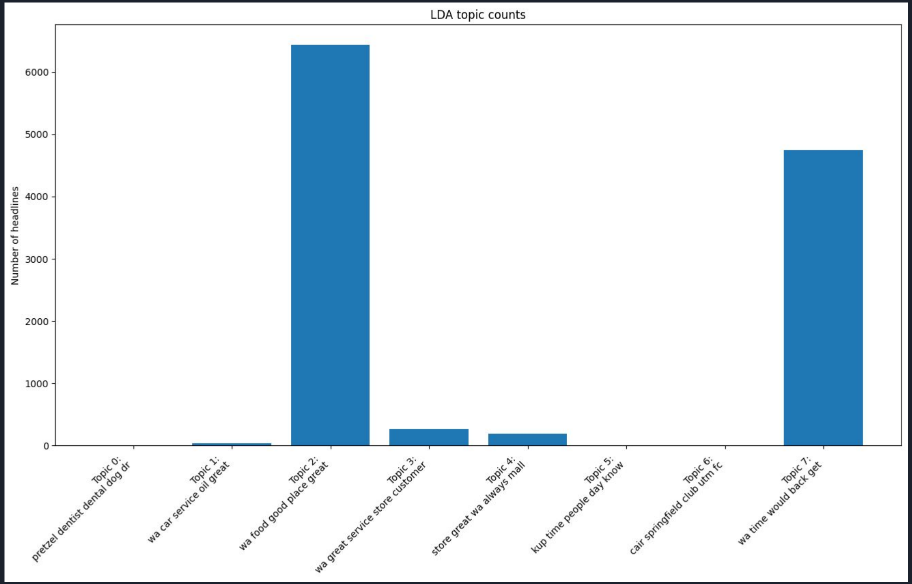
Vader Results



Modelling Results

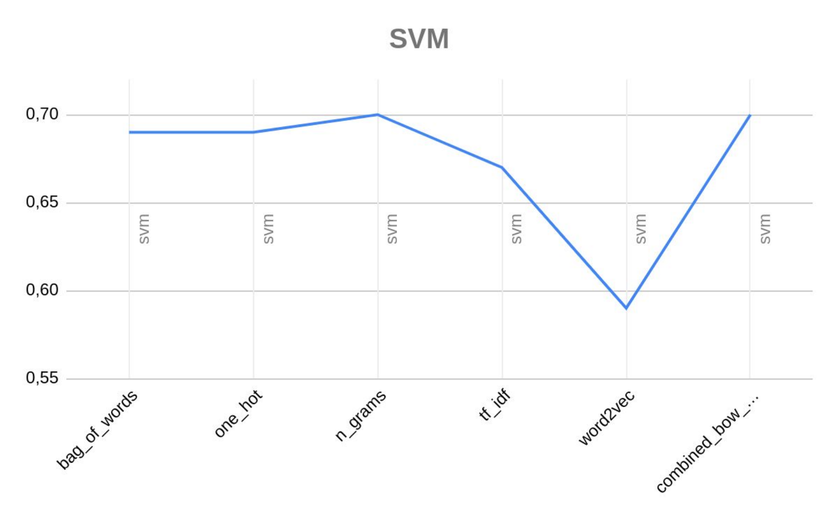
Topic Modelling

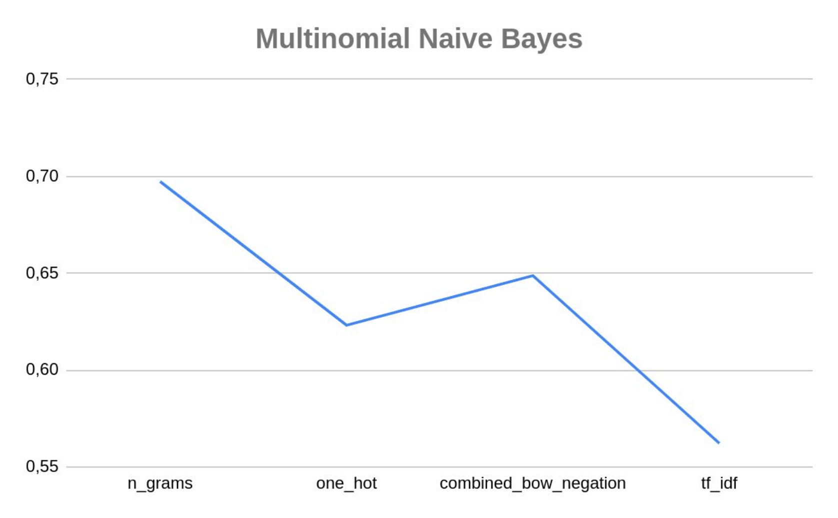




Modelling Results

Modelling





NLP - Modelling Results

Classification Results

Feature_set	Classifier	Precision	Recall	F1_score	Accuracy
bag_of_words	decision_tree	0.56	0.53	0.51	0.69
one_hot	decision_tree	0.59	0.53	0.52	0.69
n_grams	decision_tree	0.58	0.53	0.51	0.69
tf_idf	decision_tree	0.6	0.54	0.52	0.69
word2vec	decision_tree	0.43	0.42	0.41	0.56
combined_bow_negation	decision_tree	0.61	0.58	0.58	0.75
bag_of_words	logistic_regression	0.69	0.67	0.67	0.81
one_hot	logistic_regression	0.72	0.68	0.69	0.82
n_grams	logistic_regression	0.71	0.68	0.68	0.82
tf_idf	logistic_regression	0.73	0.64	0.62	0.83
word2vec	logistic_regression	0.6	0.59	0.59	0.72
combined_bow_negation	logistic_regression	0.7	0.67	0.68	0.82
bag_of_words	multinomial_nb	0.66	0.64	0.64	0.8
one_hot	multinomial_nb	0.65	0.63	0.62	0.8
n_grams	multinomial_nb	0.7	0.71	0.7	0.78
tf_idf	multinomial_nb	0.54	0.59	0.56	0.79
combined_bow_negation	multinomial_nb	0.66	0.65	0.65	0.81
bag_of_words	perceptron	0.67	0.66	0.66	0.79
one_hot	perceptron	0.68	0.66	0.67	0.8
n_grams	perceptron	0.7	0.68	0.69	0.81
tf_idf	perceptron	0.69	0.65	0.65	0.82
word2vec	perceptron	0.58	0.57	0.57	0.7
combined_bow_negation	perceptron	0.67	0.67	0.67	0.79
bag_of_words	random_forest	0.79	0.61	0.58	0.8
one_hot	random_forest	0.79	0.6	0.57	0.8
n_grams	random_forest	0.79	0.61	0.58	0.81
tf_idf	random_forest	0.79	0.6	0.57	0.8
word2vec	random_forest	0.67	0.49	0.46	0.66
combined_bow_negation	random_forest	0.79	0.61	0.58	0.81
bag_of_words	svm	0.71	0.61	0.58	0.81
one_hot	svm	0.75	0.62	0.6	0.82
n_grams	svm	0.78	0.61	0.58	0.81
tf_idf	svm	0.8	0.64	0.62	0.83
word2vec	svm	0.84	0.57	0.54	0.76
combined_bow_negation	svm	0.66	0.61	0.58	0.8
bag_of_words	xgb	0.74	0.65	0.65	0.81
one_hot	xgb	0.73	0.64	0.65	0.81
n_grams	xgb	0.73	0.65	0.65	0.81
tf_idf	xgb	0.74	0.65	0.65	0.81
word2vec	xgb	0.61	0.53	0.51	0.72
combined_bow_negation	xgb	0.72	0.64	0.64	0.81

TS - Modelling Results

Negative Reviews

	Model	MAE	MSE	RMSE	MAPE	MASE
0	baseline_mean_count	57.396113	5503.970664	74.188750	47.668328	3.550275
1	baseline_median_count	66.063910	7212.167293	84.924480	54.057759	4.086427
2	baseline_naive_count	17.030075	530.308271	23.028423	20.478147	1.053407
3	ses_count	16.663242	546.559441	23.378611	19.256359	1.030716
4	holt_linear_damped_count	16.862159	557.816911	23.618148	19.933698	1.043020
5	box_cox_arima_count	18.526316	653.037594	25.554600	21.356833	1.145958
6	arima_count	17.111485	562.465370	23.716352	20.901066	1.058442

Modelling Results

Neutral Reviews

	Model	MAE	MSE	RMSE	MAPE	MASE
0	baseline_mean_count	19.270736	661.708332	25.723692	38.658690	2.245134
1	baseline_median_count	21.390977	827.375940	28.764143	40.882695	2.492153
2	baseline_naive_count	8.992481	130.766917	11.435336	24.900605	1.047668
3	ses_count	8.335730	123.688874	11.121550	23.938086	0.971153
4	holt_linear_damped_count	8.347448	125.880421	11.219644	23.999746	0.972518
5	box_cox_arima_count	8.375940	124.601504	11.162504	23.659034	0.975838
6	arima_count	8.936647	138.417251	11.765086	25.911462	1.041163

Modelling Results

Positive Reviews

	Model	MAE	MSE	RMSE	MAPE	MASE
0	baseline_mean_count	163.481013	48932.673871	221.207310	44.069199	4.311587
1	baseline_median_count	192.078947	66887.934211	258.627018	50.408687	5.065818
2	baseline_naive_count	39.954887	3478.601504	58.979670	15.281865	1.053755
3	ses_count	40.583705	3648.026880	60.398898	15.634892	1.070339
4	holt_linear_damped_count	41.587326	3768.983277	61.392046	16.212750	1.096809
5	box_cox_arima_count	45.278195	4306.571429	65.624473	18.526608	1.194150
6	arima_count	45.347134	4395.832203	66.301072	18.687406	1.195968

Modelling Results

Sentiment /month

	Model	MAE	MSE	RMSE	MAPE	MASE
0	baseline_mean_sentiment	0.070857	0.008519	0.092298	20.466051	0.943320
1	baseline_median_sentiment	0.071587	0.008750	0.093541	20.790576	0.953036
2	baseline_naive_sentiment	0.070159	0.009187	0.095850	18.465155	0.934023
3	ses_sentiment	0.060078	0.006458	0.080365	16.358471	0.799821
4	holt_linear_damped_sentiment	0.066850	0.007259	0.085202	17.779440	0.889980
5	box_cox_arima_sentiment	0.436283	0.199034	0.446132	102.534295	5.808245
6	arima_sentiment	0.058748	0.006224	0.078894	16.014116	0.782116

Modelling Results

Restaurants Sentiment /month

	Model	MAE	MSE	RMSE	MAPE	MASE
0	baseline_mean_sentiment	0.065791	0.007682	0.087646	16.086306	0.801780
1	baseline_median_sentiment	0.066293	0.007953	0.089178	16.384731	0.807899
2	baseline_naive_sentiment	0.074367	0.010058	0.100290	17.037438	0.906290
3	ses_sentiment	0.058400	0.006160	0.078483	13.851939	0.711715
4	holt_linear_damped_sentiment	0.062746	0.006625	0.081396	14.624643	0.764677
5	box_cox_arima_sentiment	0.474841	0.232242	0.481915	103.259700	5.786790
6	arima_sentiment	0.063945	0.007058	0.084012	15.290485	0.779287

Modelling Results

Nightlife Sentiment /month

	Model	MAE	MSE	RMSE	MAPE	MASE
0	baseline_mean_sentiment	0.094769	0.016124	0.126981	24.212041	0.813998
1	baseline_median_sentiment	0.094216	0.016322	0.127757	24.746331	0.809246
2	baseline_naive_sentiment	0.115686	0.024022	0.154989	25.847463	0.993656
3	ses_sentiment	0.092929	0.016544	0.128622	23.062401	0.798193
4	holt_linear_damped_sentiment	0.095229	0.017024	0.130475	22.192033	0.817945
5	box_cox_arima_sentiment	0.479851	0.245280	0.495258	99.688450	4.121572
6	arima_sentiment	0.096136	0.018906	0.137498	21.548300	0.825742

Modelling Results

Breakfast & Brunch Sentiment /month

	Model	MAE	MSE	RMSE	MAPE	MASE
0	baseline_mean_sentiment	0.112873	0.019940	0.141208	29.690248	0.772863
1	baseline_median_sentiment	0.114420	0.020197	0.142116	30.220162	0.783450
2	baseline_naive_sentiment	0.137241	0.030459	0.174525	35.022399	0.939714
3	ses_sentiment	0.107631	0.018532	0.136134	27.241732	0.736967
4	holt_linear_damped_sentiment	0.111707	0.020183	0.142068	27.925589	0.764878
5	box_cox_arima_sentiment	0.461476	0.231206	0.480838	105.552306	3.159803
6	arima_sentiment	0.105781	0.017907	0.133818	26.880647	0.724299