PCE estimation using time series

Business understanding

The provided dataset is seasonally adjusted, which can be used for time series analysis and estimation. There is an unusual fluctuation in data regarding the recent pandemic that needs to be handled before the analysis. This project aims to analyse the dataset in terms of time series and estimate the personal consumption expenditures (PCE) for October 2022. We will conduct CRISP-DM methodology in our analysis.

Data understanding

The data is seasonally adjusted. That means the effect of seasonal variations is removed from time series. A seasonally adjusted dataset still contain ***remainder*** and ***trend-cycle*** component, so we need to decompose these particles. Seasonally adjusted time series are not "***smoothed***" and still need a decomposition to smooth the up and downtrends. (Hyndman and Athanasopoulos, 2018)

We decided to limit the time window of the analysis. The dataset contains all possible data from 1950 till Dec. 2021. Using old data does not necessarily improve the model performance and may mislead the model as the old trends and seasonal components may have changed over time. So we limited the analysis time frame to the window of 1/1990 to 12/2021. (about 380 observations)

There are 8 ***NA*** records in the PCE column that need to be imputed by a suitable amount. There are different options for imputation, but we use a time series based model for this purpose.

Chart 1: PCE trend

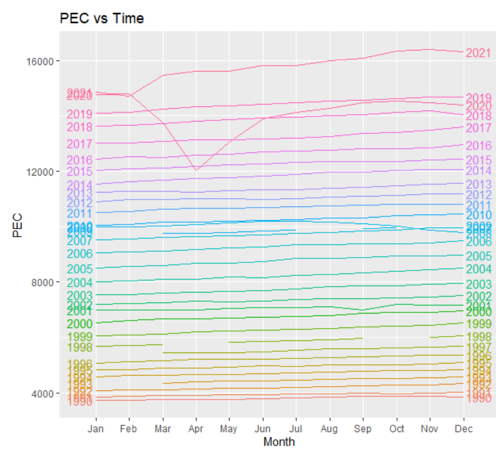


Chart 1 demonstrates that the seasonal fluctuations of the dataset were removed ideally from the data, but the dataset is not smoothed yet. Testing the stationary assumption using the "***Augmented Dickey-Fuller***" Test in R shows that the time series is not stationary. We will conduct the differences to make it stationary.

```
        Augmented Dickey-Fuller Test

data:  imputed
Dickey-Fuller = -1.4638, Lag order = 6, p-value = 0.8017
alternative hypothesis: stationary
```

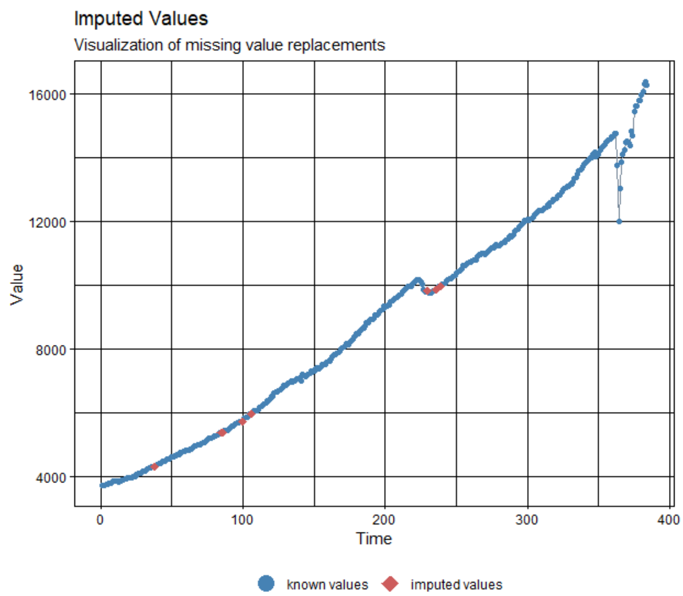
Fig 1: Stationary test result

Classical smoothing models like moving averages are not robust to unusual outliers, So we have to use more robust models like `X11` or `STL` for smoothing.

Data preparation

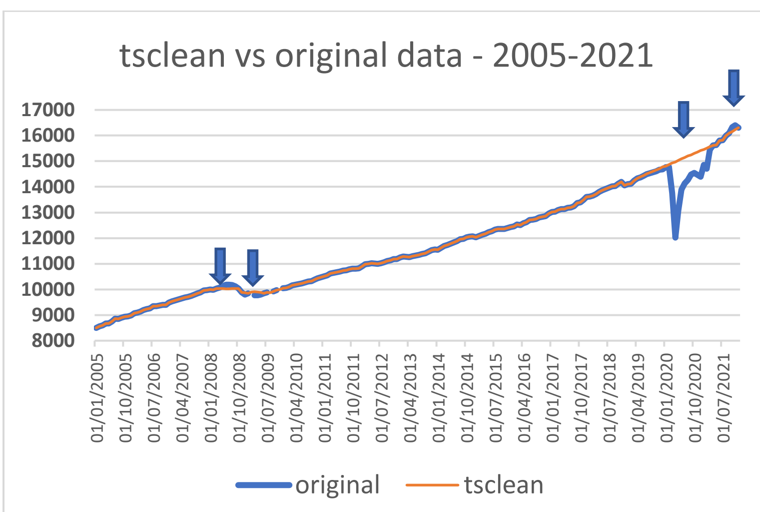
We have imputed the missing data using `na_kalman` package in R. The package uses `auto.ARIMA` model to impute the missing data.

Chart 1: Imputed data



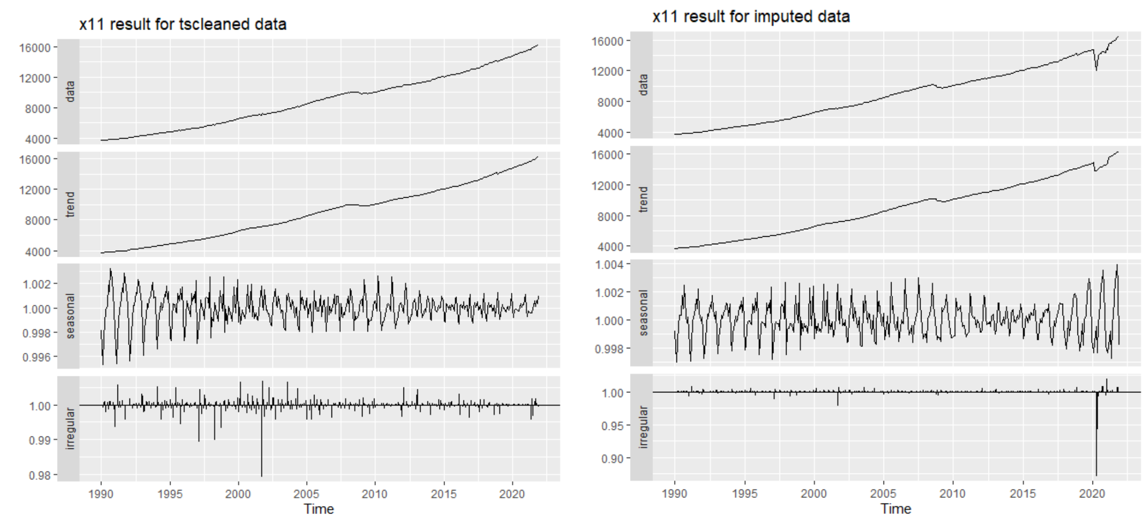
Moreover, we can see a set of outliers in the plot. ***ARIMA*** Model modify itself by the "***Theta"*** component to cover the outliers, but there is no such modification in other models. We use `tsclean` package that performs "***Friedman's super smoother***" for non-seasonal series and `STL` model for seasonal series to fill the missing data and outlier replacement. The model defines an upper(***U***), and lower(***L***) bound for residuals and replaces outliers with either linear ***interpolation*** or ***STL*** fit. Running the model results in a smoothed dataset without outliers and noises:

Chart 2: tsclean trend vs original data



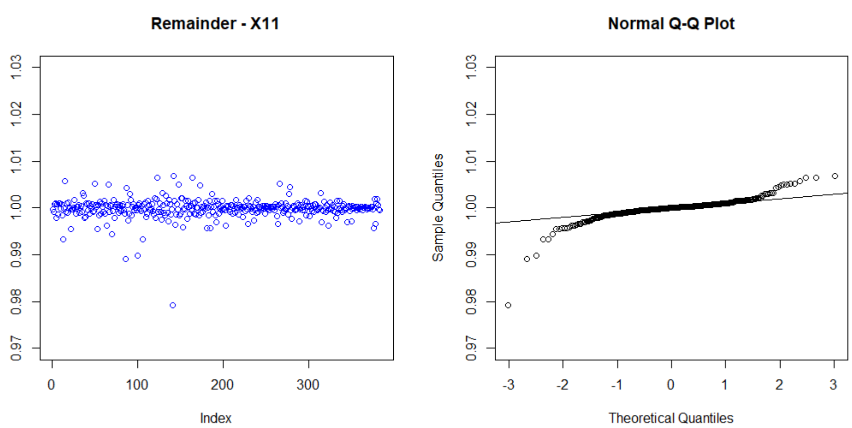
Testing different decomposer models, including ***STL*** and ***MA***, we achieved the best performance using ***X11*** model. The model is robust to outliers and perfectly handles the shock. We conducted the `SEAS` function, and it showed an excellent decomposition performance. Most of the data is described by the trend and seasonal pattern in original and cleaned data.

Chart 3: Decomposed time series



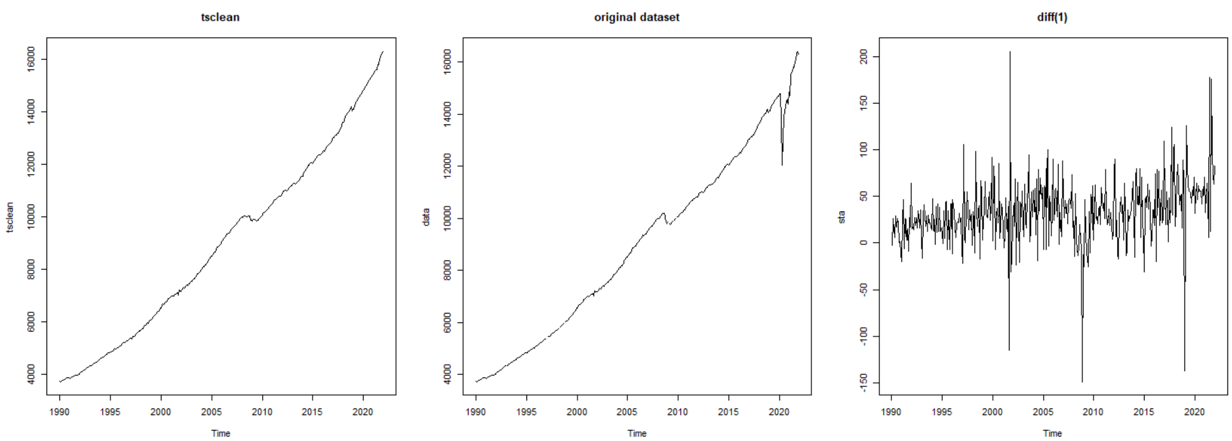
Despite the seasonal adjustment, we have some seasonal components which are not significant. The remainder seems noisy and is empty of a pattern. The cleaned data resulted in better data capturing in trend and seasonal components. Also, we checked if remainders are ***normally distributed*** using the Q-Q plot from `qqnorm` package, and the result is acceptable.

Chart 4: Remainder distribution check



ARIMA model makes the series stationery itself, but we used the `diff` function to alternate the first difference records to make the series stationary for other models. Testing the stationary demonstrates a good result which is meaningful for at 5% level.

Chart 5: data preparation results

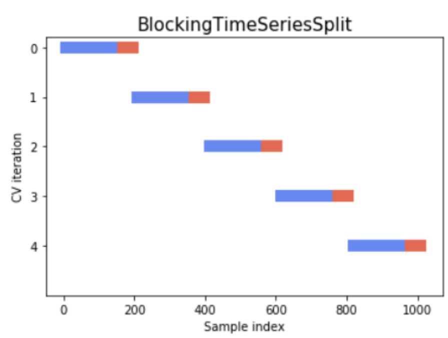



```
        Augmented Dickey-Fuller Test

data:  sta
Dickey-Fuller = -7.3094, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary
```

Fig 2:Stationary test results after differentiating

We partitioned the cleaned and original dataset into six-folds and added to one whole dataset to cross-validate (***blocked cross-validation***) the models' accuracy performance without reestimating the parameters. Blocked cross-validation prevents the model from information leakage from the future and avoids overfitting the model. (Pirbazari et al., 2021) We have 60 training and 15 test observation for each fold.

Chart 6: demonstration of blocking cross-validation




```
P1ITR <- window(imp1, start =c(1990,1), end =c(1994,12))#part1 imputed training
P1ITE <- window(imp1, start =c(1995,1), end =c(1996,4))#part1 imputed test

P2ITR <- window(imp1, start =c(1994,2), end =c(1999,1))#part2 imputed training
P2ITE <- window(imp1, start =c(1999,2), end =c(2000,5))#part2 imputed test

P3ITR <- window(imp1, start =c(1999,4), end =c(2004,3))#part3 imputed training
P3ITE <- window(imp1, start =c(2004,4), end =c(2005,7))#part3 imputed test

P4ITR <- window(imp1, start =c(2004,8), end =c(2009,7))#part4 imputed training
P4ITE <- window(imp1, start =c(2009,8), end =c(2010,11))#part4 imputed test

P5ITR <- window(imp1, start =c(2010,2), end =c(2015,1))#part5 imputed training
P5ITE <- window(imp1, start =c(2015,2), end =c(2016,5))#part5 imputed test

P6ITR <- window(imp1, start =c(2015,9), end =c(2020,8))#part6 imputed training
P6ITE <- window(imp1, start =c(2020,9), end =c(2021,12))#part6 imputed test
```

Fig 2: Partitions

Retaining or removing the outliers from a series is a challenging decision. Outliers in our data are actual observations that resulted from the pandemic; simultaneously, we know these are not the usual fluctuation of our time series. As (Hyndman and Athanasopoulos, 2018) suggested, we decided to use both data series for our analysis. This project's first aim is to compare the prediction models, so we would evaluate the models for both datasets.

Modelling

We have conducted the "***Random walk with drift model***", "***simple exponential model***", and "***ARIMA***" model to analyse both datasets. We defined "***short=3 period***" and "***long term=16 period***" evaluation periods to have a fair comparison between models. We run the models for each partition and each period window, then evaluate the model's accuracy using the average accuracy measures for both original and cleaned data.(just focused on error on test set not training)

Random Walk Drift model

We used `rwf` package and set the ***drift*** option to true to let the function use drift to capture the time series trend. We tested the model for all data (Fig. 3). The drift for this model is 32.6237. We will discuss the result for different folds later.

```
Forecast method: Random walk with drift

Model Information:
Call: rwf(y = pallTTR, h = h, drift = TRUE)

Drift: 32.6237  (se 1.7731)
Residual sd: 34.6098

Error measures:
                       ME     RMSE      MAE         MPE      MAPE       MASE        ACF1
Training set 2.244027e-13 34.56439 24.28733 -0.05072038 0.3004614 0.06170768 -0.01951767
```

Fig 3: rwf with drift

Simple exponential

We set the initial to ***optimal*** to let the `ses` uses `ets` select the best trend, seasonal, and error parameters and tune the parameters.

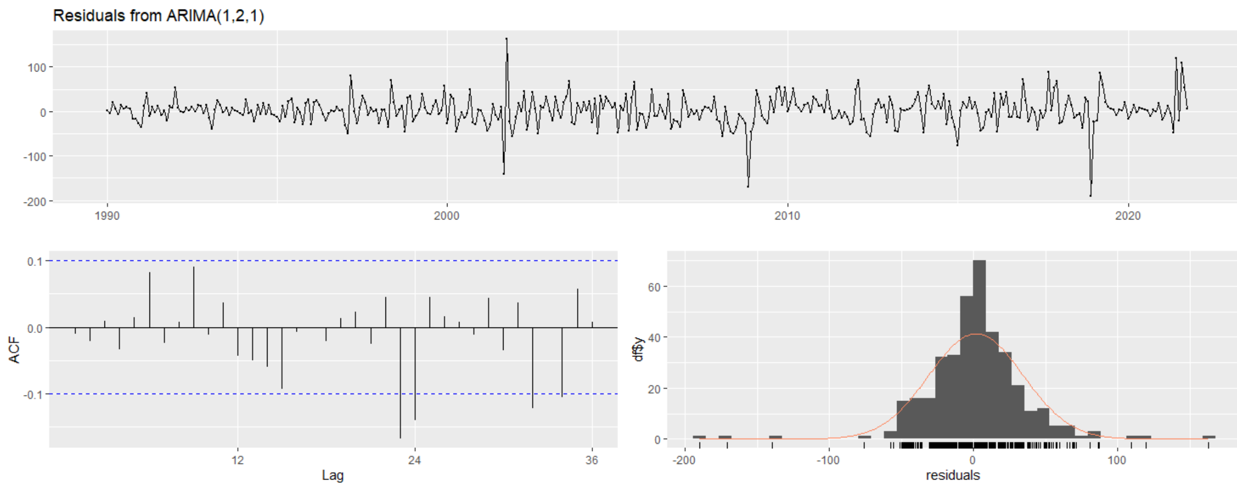
```
Forecast method: Simple exponential smoothing

Model Information:
Simple exponential smoothing

Call:
 ses(y = pallTTR, h = h, initial = "optimal")

  Smoothing parameters:
    alpha = 0.9999

  Initial states:
    l = 3730.5353

  sigma:  47.5937

     AIC     AICc      BIC
5226.248 5226.312 5238.084

Error measures:
                 ME     RMSE      MAE       MPE      MAPE       MASE      ACF1
Training set 32.542 47.46893 37.55544 0.3822721 0.4436485 0.09541841 -0.016864
```

Fig 4: Simple exponential forecaster

Arima model

We used `auto.arima` to optimise the ARIMA model for the dataset. We ran the auto arima model for the whole dataset and then used the optimised, fixed parameters for other folds to fairly judge the models. ***<u>If we let the models reestimate the parameters in each fold, it would not be possible to compare other models</u>***. So we fix the parameters as ARIMA(1,2,1). The function selected a ***damped-trend linear, exponential smoothing*** model.

Chart 7: ARIMA(1,2,1) performance




```
Series: pallTTR
ARIMA(1,2,1)

Coefficients:
          ar1      ma1
      -0.2092  -0.8958
s.e.   0.0558   0.0311

sigma^2 = 1072:  log likelihood = -1864.94
AIC=3735.89   AICc=3735.95   BIC=3747.71
```

```
        Ljung-Box test

data:  Residuals from ARIMA(1,2,1)
Q* = 35.141, df = 22, p-value = 0.03745

Model df: 2.   Total lags used: 24
```

Fig 5: Auto-ARIMA selected model

The autocorrelation plot has some minor outlier amounts, and the p-value rejects the null hypothesis at the 3%-level. We use ARIMA(1,2,1) (1 step autoregressive, second-order differencing) for all the analyses afterwards in this report. The models' prediction vs actual for 16-periods in the last three folds plotted for both original and cleaned data are as chart8:

Chart 8: prediction of all models for h=16 and cleaned data

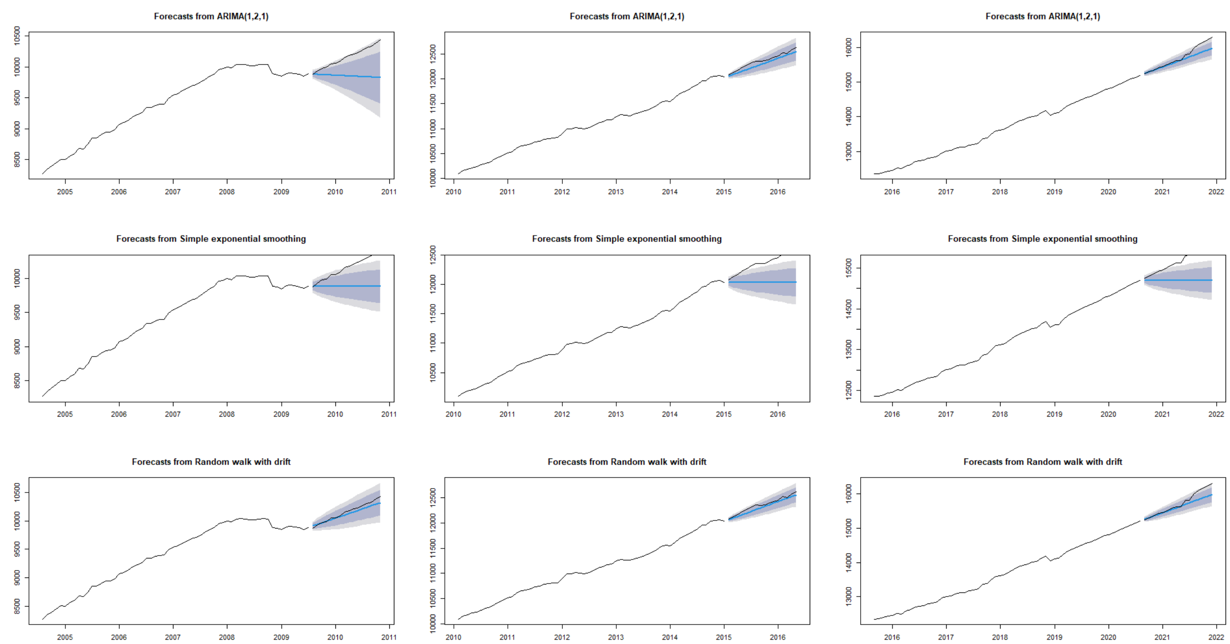
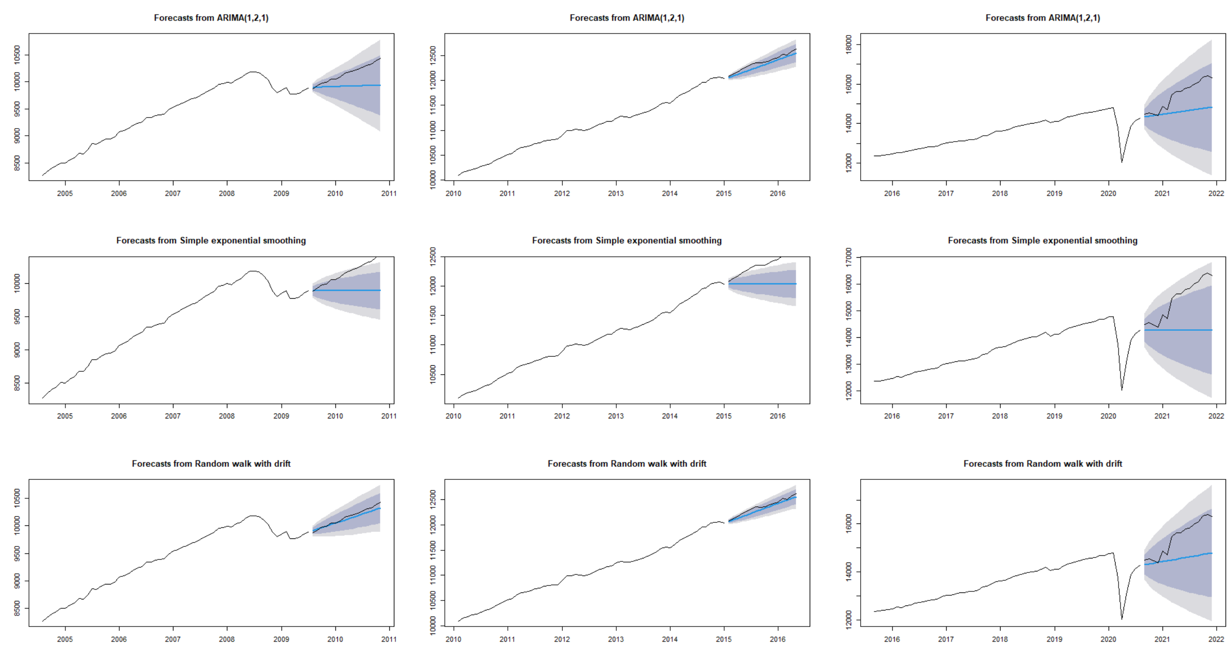


Chart 9: prediction of all models for h=16 and original data



RWF drift has performed better in some folds and is as good as the ARIMA model in others.

Evaluation

We investigated a common dataset for different models so that we could use each MPE, MAPE, MASE and RMSE error evaluation indices. These indices are the most suitable in the problem context. We calculated the ***average absolute amount*** of each error index for different folds and then compared the models' performance using the majority role to select the best model. The dominant model is the best one.

FOR h=16 (cleaned)	RMSE	MPE	MAPE	MASE
Sim. Exponential	382.5473	3.51497	3.517142	0.877669
Drift	92.64985	0.741043	0.813393	0.198532
ARIMA(1,2,1)	140.0628	1.168248	1.199095	0.302858
best model supremacy	33.85%	36.57%	32.17%	34.45%

Table 1: Error indices for h=16-cleaned

For h=16 and cleaned-data, rwf model dominates in all error evaluation indices. We can firmly select it as the best model in this setting. The supremacy (less average error) of the best model over the second-best model is 35%. Also, we investigated the original data.

FOR h=16 (original)	RMSE	MPE	MAPE	MASE
Sim. Exponential	507.1937	4.194674	4.196846	1.039821
Drift	241.3933	1.559559	1.630932	0.408231
ARIMA(1,2,1)	273.3984	1.85204	1.889527	0.472898
best model supremacy	11.71%	15.79%	13.69%	13.67%

Table 2: Error indices for h=16-original

The supremacy of the rwf-drift decreased by using original data. That means the rwf is not robust against the fluctuations in the time series, and generally, ARIMA models update faster in these situations. However, overall the random walk drift demonstrated better performance in both scenarios.

By setting the h=2, ARIMA(1,2,1) shows better performance on original and cleaned data.

	FOR h=2	RMSE	MPE	MAPE	MASE
tscleaned	Sim. Exponential	55.2667	0.6420	0.5420	0.0732
	Drift	32.5846	0.5315	0.2456	0.0550
	ARIMA(1,2,1)	27.5495	0.0203	0.1965	0.0346
	best model supremacy	15.45%	96.17%	20.01%	37.15%
original data	Sim. Exponential	199.3320	0.9547	0.8126	0.1786
	Drift	71.3820	0.7217	0.4933	0.1516
	ARIMA(1,2,1)	57.6815	0.3843	0.4057	0.1235
	best model supremacy	19.19%	46.75%	17.75%	18.56%

Table 3: Error indices for h=2-original

One step ahead, rolling up

To complete our evaluation, we conducted the one step ahead roll up to cross-validate the models for h=1. The result for the first 80% training and 20% test dataset is as below:

	RMSE	MPE	MAPE
Drift	25.40553	0.437968	0.492319
ARIMA(1,2,1)	16.67721	0.038875	0.305536
Sim. Exponential	16.50887	0.035622	0.298359

Table 4: Error indices for h=1-original

The simple exponential smoothing performs better than rwf and ARIMA models for one step ahead prediction. Despite its simple structure, this model performs very efficiently for one step ahead prediction. The supremacy of this model is very low compared to the ARIMA model.

Finally, we can conclude that the RWF model performs better according to our six folds results in longer horizon prediction on both cleaned and original data, but ARIMA (1,2,1) predicts better in the shorter windows. A simple exponential performs slightly better than ARIMA for one-step ahead prediction. Hence, we use the RWF model on the original time series to predict PCE for October 2022.

Deployment

Running the rwf-drift model on the whole data set and for the original dataset predicts the PCE amount for October 2022, 16634.64 with a 95% confidence interval of 15765.53 and 17503.76. Running the model for cleaned data has a very similar prediction with a tighter confident interval.

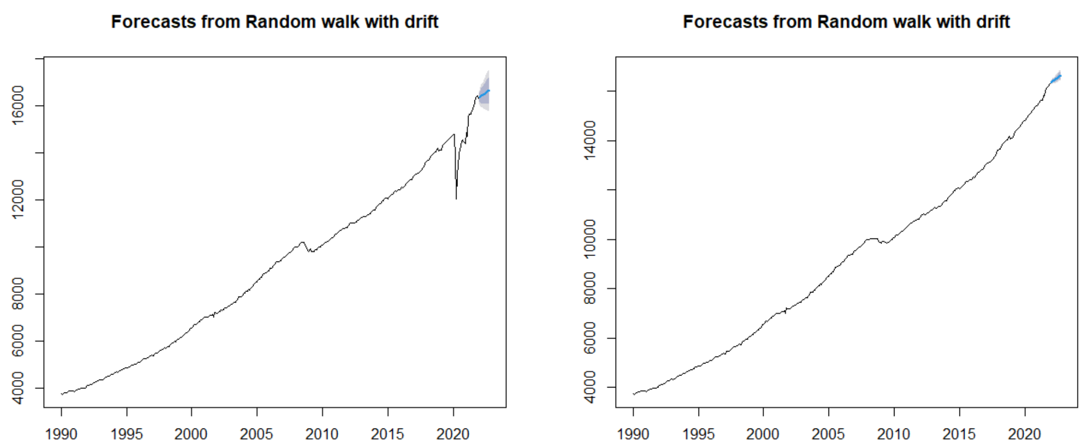
	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan 2022	16339.13	16161.50	16516.77	16067.47	16610.80
Feb 2022	16371.97	16120.43	16623.51	15987.28	16756.66
Mar 2022	16404.80	16096.33	16713.27	15933.04	16876.57
Apr 2022	16437.64	16080.99	16794.29	15892.19	16983.09
May 2022	16470.47	16071.21	16869.74	15859.85	17081.10
Jun 2022	16503.31	16065.37	16941.24	15833.54	17173.07
Jul 2022	16536.14	16062.51	17009.78	15811.78	17260.50
Aug 2022	16568.98	16061.99	17075.96	15793.60	17344.35
Sep 2022	16601.81	16063.38	17140.24	15778.35	17425.27
Oct 2022	16634.64	16066.36	17202.93	15765.53	17503.76

Fig 6: RWF forecast for upcoming months - original

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan 2022	16339.13	16294.73	16383.54	16271.22	16407.05
Feb 2022	16371.97	16309.09	16434.85	16275.80	16468.13
Mar 2022	16404.80	16327.69	16481.92	16286.87	16522.74
Apr 2022	16437.64	16348.48	16526.79	16301.28	16573.99
May 2022	16470.47	16370.66	16570.28	16317.83	16623.12
Jun 2022	16503.31	16393.83	16612.78	16335.88	16670.74
Jul 2022	16536.14	16417.74	16654.54	16355.06	16717.22
Aug 2022	16568.98	16442.24	16695.71	16375.15	16762.80
Sep 2022	16601.81	16467.21	16736.41	16395.96	16807.66
Oct 2022	16634.64	16492.58	16776.70	16417.38	16851.91

Fig 7: RWF forecast for upcoming months - cleaned

Chart 10: rwf prediction



Section 2: Topic modelling of amazon comments

Intro

Businesses use ***topic modelling*** to extract their customers' attitudes to their products from their comments on different platforms. This part of the report is devoted to extracting the most important topics of the 5000 comments from the Amazon website. The data is diverse, and the comments are about different brands and models. So, the final result of our study may not be helpful in terms of a specific product or brand analysis.

Business and data understanding

The data has an instrumental variable: the product score out of 5. We assume this variable as a Likert measure (1-5) for satisfaction. Higher stars mean more satisfaction and vice versa. This project aims to topic modelling of positive and negative comments. So, we can divide the comments into positive(P) and negative(N) according to the users' starts given to the products.

We assume the satisfied customers have given 4 or 5-stars to products and unsatisfied users have given 1 or 2-stars. The comments with 3-star can not be counted as positive or negative comments. There are other options like defining more than two satisfaction classes for comments, but we focused on two classes for comments as it is asked. Also, other possible factors like the colour and size_name can be processed to measure the customer's satisfaction, which is off the comment analysis topic. We will use both ***titles*** and ***comment*** fields text for our analysis. ***Titles*** are usually more concise and use fewer general words.

It is worth mentioning that another solution for comment dividing is text sentiment analysis using Tidyverse `get_sentiments("bing")` ***function. As we have access to helpful product score data, using sentiment analysis does not make sense. As (Al-Natour and Turetken, 2020) suggested, at this time, sentiment analysis methods can be used as a complementary factor but not a perfect substitute where ratings exist.***

Data preparation

We need to ensure the format of the text column. We used `str_conv` to convert the strings into UTF-8. Also, we defined two classes of "***satisfied***" and "***unsatisfied***" customers regarding stars. The "***satisfied***" group has 3729 members, and the "***unsatisfied***" has 626 members. From this point afterwards, we analyse these two classes separately.

We lemmatised the documents using the `lemmatize_string` function, then tokenised the documents by `tokeniser` to be able to process them. We converted the combination of "***titles***" and "***comments***" to a corpus document using the `corpus` function. The next step is producing the document term matrix(DTM). `DocumentTermMatrix` function has a lemmatisation sub-function as a part of its controls. Setting the controls to remove punctuations, numbers, and stop words, we removed words with less than one character and lowered all characters.

Furthermore, we tested both the ***TF*** and ***TF_IDF*** methods. The aim of this project is topic extraction, and also the comments are usually free of prefixes and common words. So, reducing the weights of the common words among the comments could lose some critical tokens. Hence, we decided to use ***TF*** instead of the ***TF-IDF*** method. ***TF-IDF*** gives more weight to the rare words with less frequency among the documents. The frequent terms of the ***TF-IDF*** model are very unspecific in this case.

```
> findFreqTerms(dtmpo,lowfreq = 300)
  [1] "days"       "delivery"   "fast"        "good"      "money"    "using"    "value"     "backup"  "battery"    "best"
 [11] "better"     "camera"     "like"        "phone"     "screen"   "time"     "well"      "work"    "budget"     "nice"
 [21] "product"    "experience" "gaming"      "low"       "price"    "really"   "awesome"   "range"   "smartphone" "excellent"
 [31] "overall"    "looks"      "nord"        "review"    "will"     "great"    "processor" "just"    "k"          "one"
 [41] "plus"       "premium"    "worth"       "mobile"    "note"     "redmi"    "issue"     "super"   "quality"    "device"
 [51] "everything" "perfect"    "buy"         "finger"    "working"  "charging" "speed"     "pro"     "amazing"    "mi"
 [61] "superb"     "amazon"     "light"       "first"     "oneplus"  "average"  "back"      "front"   "life"       "performance"
 [71] "phones"     "expected"   "except"      "can"       "day"      "bit"      "satisfied" "ok"      "look"       "colour"
 [81] "display"    "also"       "fingerprint" "loved"     "go"       "must"     "problem"   "smooth"  "reader"     "much"
 [91] "months"     "bad"        "features"    "gb"        "u"        "print"    "dont"      "issues"  "design"     "little"
[101] "looking"    "bought"     "decent"      "segment"   "love"     "happy"    "thanks"    "now"     "got"        "purchase"
[111] "use"
> findFreqTerms(dtmpo2,lowfreq = 300)
 [1] "good"    "money"   "battery" "best"    "camera"  "phone"   "nice"    "product" "price"   "awesome" "great"   "one"     "mobile"  "quality"
```

Fig 1: Frequent words–TF vs TF-IDF

Constructing the DTM, we can see 3728 documents(rows) and 4191 terms in the matrix. About 99.8% of all entries are sparse and need modification. For the negative comments, this number is 99.4%.

P-comments:

```
<<DocumentTermMatrix (documents: 3728, terms: 4191)>>
Non-/sparse entries: 31413/15592635
Sparsity           : 100%
Maximal term length: 43
Weighting          : term frequency (tf)
```

Fig 2: P-DTM

N-comments:

```
<<DocumentTermMatrix (documents: 624, terms: 1873)>>
Non-/sparse entries: 6778/1161974
Sparsity           : 99%
Maximal term length: 20
Weighting          : term frequency (tf)
```

Fig 3: N-DTM

By removing the sparse tokens with a trigger of 97%, we achieved a sparsity level of 91% for positive comments. The number of the remaining terms for positives is 48. For negatives, this number is 57 and 92% sparsity.

```
> findFreqTerms(dtmspo,lowfreq = 100)
 [1] "fast"      "good"      "money"       "value"      "battery"  "best"        "better"   "camera"   "like"
[10] "phone"     "budget"    "nice"        "product"    "price"    "really"      "awesome"  "range"    "excellent"
[19] "overall"   "nord"      "great"       "just"       "k"        "one"         "worth"    "mobile"   "note"
[28] "redmi"     "super"     "quality"     "everything" "buy"      "amazing"     "mi"       "superb"   "amazon"
[37] "oneplus"   "life"      "performance" "display"    "also"     "fingerprint" "go"       "features" "use"
> dtmspo
<<DocumentTermMatrix (documents: 3601, terms: 45)>>
Non-/sparse entries: 14169/147876
Sparsity           : 91%
Maximal term length: 11
Weighting          : term frequency (tf)
```

Fig 4: P-DTM removed sparse

```
> findFreqTerms(dtmsne,lowfreq = 100)
 [1] "phone"   "good"    "camera"  "quality" "buy"     "dont"    "issue"   "mobile"  "product" "battery" "bad"     "worst"
> dtmsne
<<DocumentTermMatrix (documents: 598, terms: 57)>>
Non-/sparse entries: 2571/31515
Sparsity           : 92%
Maximal term length: 13
Weighting          : term frequency (tf)
```

Fig 5: N-DTM removed sparse

Furthermore, we prepared the frequency table of the words in each document and the whole text. We used `colSum` and `rowSum` on the DTM to shape the tables that will be used by ***Latent Dirichlet Allocation (LDA)*** and ***word cloud***. The word cloud of P-terms and N-terms:

Chart 1: P-comments' terms

Chart 2: N-comments' terms

As it was predictable, the token phone and camera are the most interesting tokens in both topics. But some differentiation in most frequent words.

Topic modelling

Having the data in corpus form, we can start topic modelling. The first step is optimising the model parameter "***k***" as the number of topics. The ***k***, which makes the highest coherence score among the topics, would be a candidate for the best number. It is worth mentioning that this ***k*** is just a suggestion. We applied `CalcProbCoherence` to estimate the k.

For $k = c(2{:}15)$ and *4000* iterations, we ran a loop to measure the coherency of the ***LDA*** models. The final result of the model for P-comments suggests ***k=4*** and a ***k=7*** as the best coherent number of topics for N-comments.

Chart 3: P-comments coherence score

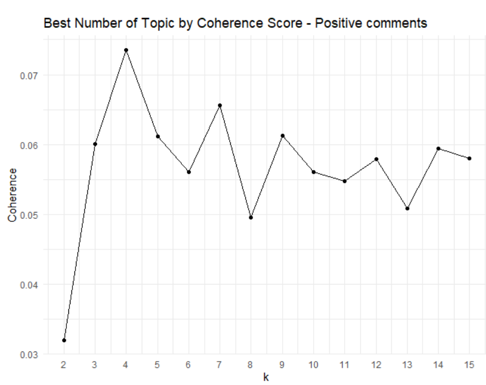
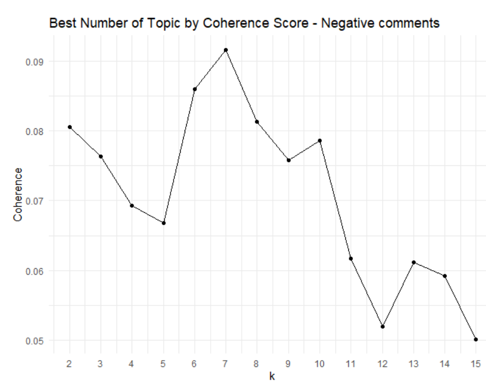


Chart 4: N-comments coherence score

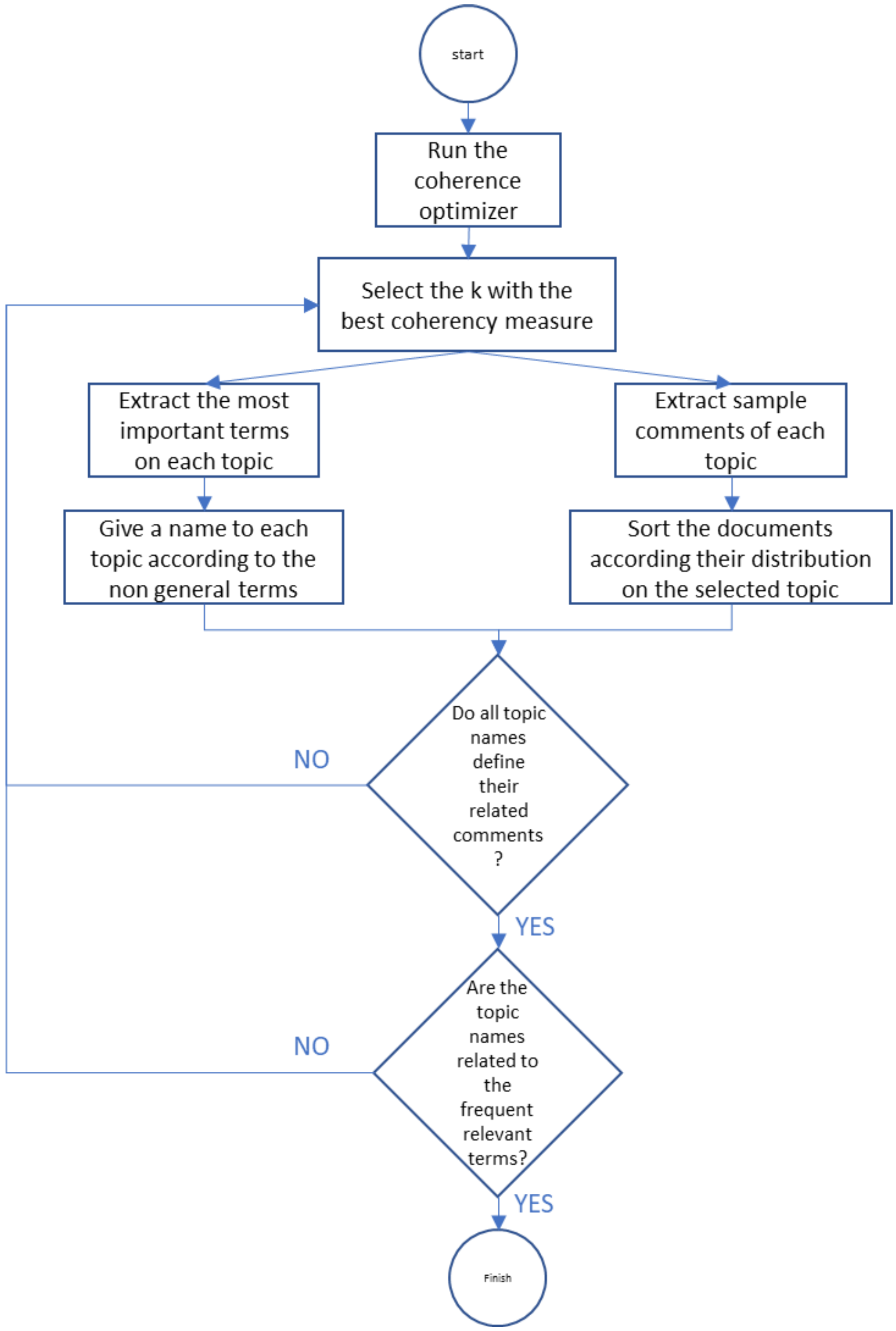


We ran the ***LDA*** model for suggested ***k***s and 4000 iterations. ***Phi*** and ***Theta*** parameters demonstrate the token distribution over the topic and distribution of documents over the topic, respectively. Investigating the ***Theta*** shows a smooth distribution for most documents over topics, which means an overlap on topics. To examine the efficiency of the suggested ***k***, we need to interpret it. So, we sampled 1000 and sorted the comments according to their distribution on each topic to find if the topics define the topic terms and comments.

Also, we have used another valuable measure, "***relevancy***", which is the result of (Sievert, 2014) study. The idea is to measure the frequency of a particular term in a topic compared to its frequency in the whole corpus. Due to the word limit of this report, we investigated the result of ***the three most popular topics (regarding the number of comments on each topic Theta)*** of positive and negative comments.

The topics must be specific, and the majority of the terms and comments(contents) must be interpretable by the topic's name. As long as we can name all the suggested topics using their contents, we can call them proper topics. If a topic cannot interpret its contents, we will repeat the process with another ***k***. Our decision model showed in Chart5. We try to extract the suitable topics according to this process.

Chart 5: k selection process



Positive comments:

We tested increasing the topics to 5, 6, and 7 and decreasing them to 3 and 2, but the results were not better.

	Topic 1	Topic 2	Topic 3	Topic 4
1	good	nice	camera	phone
2	product	mobile	battery	best
3	money	awesome	quality	price
4	value	one	good	great
5	redmi	excellent	life	budget
6	note	phone	performance	range
7	worth	like	fast	k
8	amazon	nord	also	features
9	mi	buy	display	amazing
10	overall	oneplus	fingerprint	oneplus
11	go	super	better	performance
12	k	superb	really	note
13	price	just	overall	good
14	use	everything	one	camera
15	mobile	go	use	worth

Fig 6: Most frequent terms on topics

Topic 1: ***the value of the purchase***: The most frequent words in the topic are related to purchase and evaluation of the purchase's value. The sample comments (apendix1)demonstrate that the chosen topic name is logical for the contents.

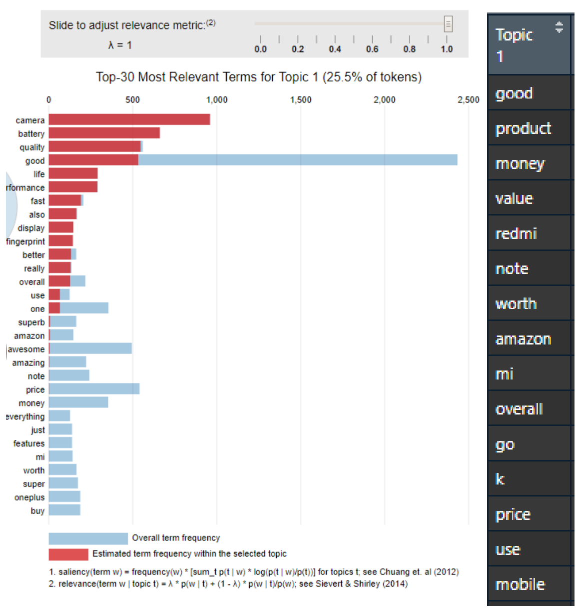


Fig 8: LDA-most relevant terms of topic 1

Topic 2: ***design and physical features***: the most frequent words on this topic and comments(appendix2) are related to phone features. Topic words "phone" and "nice" are samples of these words. However, the relevant words are slightly different from the frequent terms in this topic.

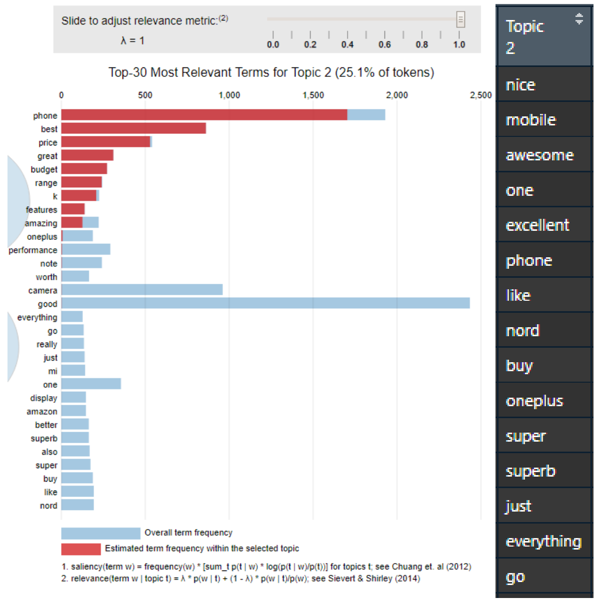


Fig 10: LDA visualisation and most relevant terms of topic 2

Topic 3*: *Phone specs and quality, including battery, display: Analysing the topic words, comments(appendix3) and most relevant words show that the comments in this topic focused on the phone specs and its quality.

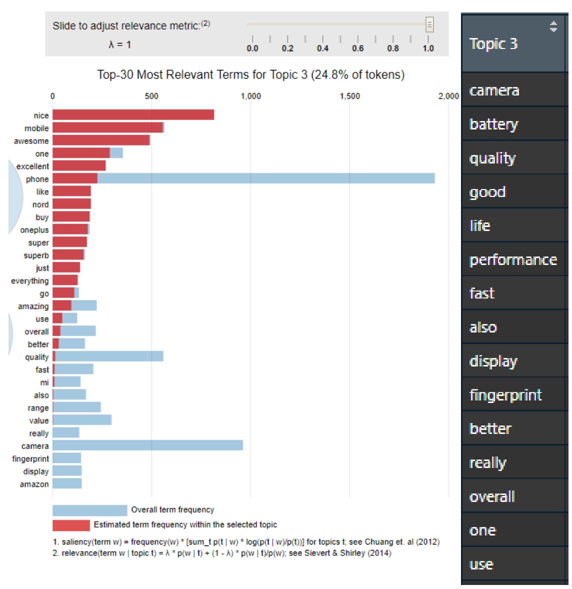


Fig 10: LDA visualisation and most relevant terms of topic 3

Topic 4*: *Product performance, benchmarking and comparison

Negative comments:

By using ***k=7***, the first five topics are interpretable, but the two remaining topics do not make sense.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
1	camera	dont	good	issue	one	phone	mobile
2	quality	buy	battery	problem	oneplus	working	bad
3	worst	product	life	display	nord	redmi	call
4	poor	money	time	network	plus	like	getting
5	got	waste	fast	heating	amazon	note	even
6	working	just	disappointed	also	using	just	average
7	screen	please	automatically	tint	service	bad	get
8	buy	plus	experience	hanging	screen	worth	performance
9	like	got	better	use	days	will	will
10	disappointed	worst	also	performance	back	mobile	purchase
11	mobile	poor	please	like	much	money	better
12	average	nord	got	experience	use	plus	worth
13	better	much	bad	purchase	please	one	time
14	much	days	just	days	automatically	hanging	screen
15	phone	worth	quality	much	time	worst	waste

Fig 11: Most frequent terms on topics for k=7 on N-comments

We tested 8,9,10, and 6, 5, and 4 topics and tried to interpret the topics. The best interpretable topic was achieved with ***k=6***. Also, on the coherence curve, the second most coherent result is achievable with ***k=6***.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
1	good	dont	phone	mobile	worst	camera
2	battery	buy	working	issue	one	bad
3	life	product	redmi	problem	nord	quality
4	amazon	money	use	display	oneplus	poor
5	screen	waste	service	network	plus	even
6	call	experience	time	heating	getting	fast
7	disappointed	performance	call	tint	using	average
8	got	worth	days	automatically	also	performance
9	hanging	plus	worst	note	just	time
10	working	better	back	worth	will	better
11	please	just	like	time	get	purchase
12	like	call	got	experience	much	money
13	days	hanging	screen	better	like	note
14	much	automatically	also	amazon	please	also
15	even	display	purchase	phone	better	waste

Fig 12: Most frequent terms on topics for k=6

Topic 1: ***performance and processor***: The most frequent terms in this topic are about product performance and the customer's personal experience using the phone.(appendix4)

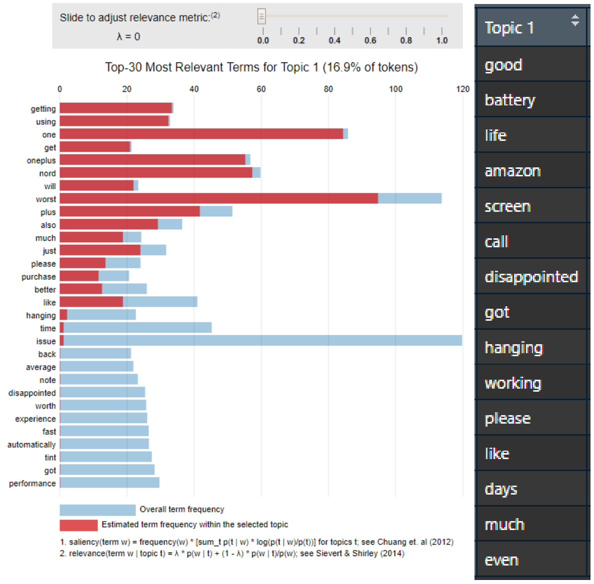


Fig 14: LDA-most relevant terms of topic 1

Topic 2: ***The value of the purchase***: the topic terms and sample comments(appendix5) show that this topic is about the purchase's value. Buyers tried to describe their general sense of their purchase and why it was not worth it.

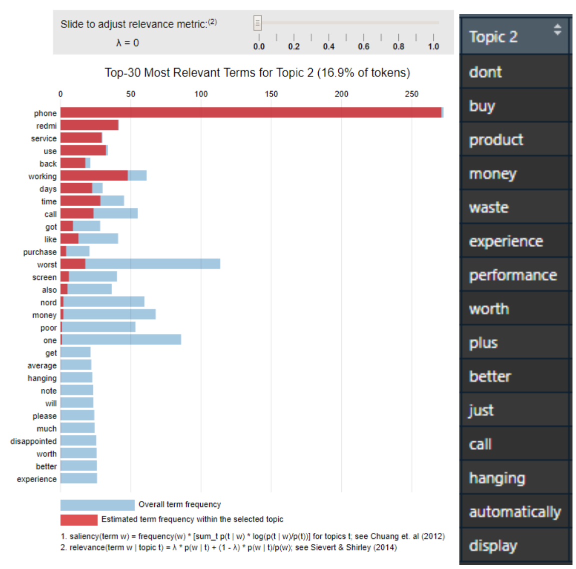


Fig 16: LDA-most relevant terms of topic 2

Topic 3: Camera and battery: The most frequent words in this topic and comments(appendix6) are related to camera and battery. Also, relevant words confirm the topic name.

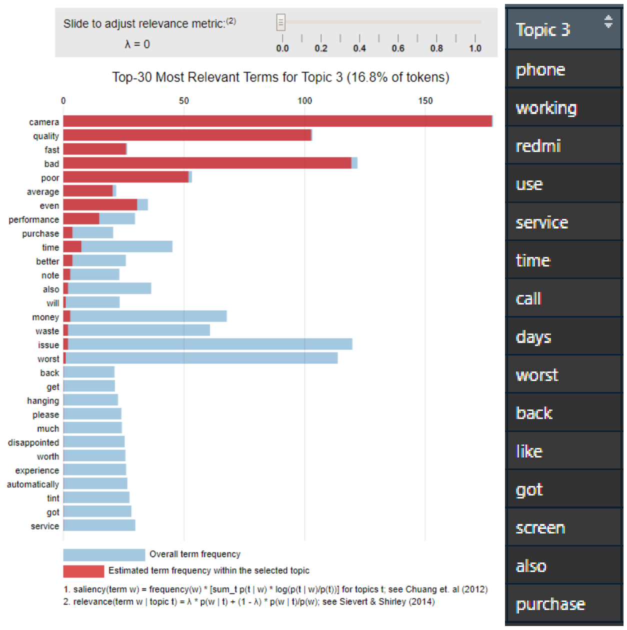


Fig 18: LDA-most relevant terms of topic 3

Topic 4: Hardware issues, including display, network and heating

Topic 5: Comparison to competitors

Topic 6: Quality of the product

References

AL-NATOUR, S. & TURETKEN, O. 2020. A comparative assessment of sentiment analysis and star ratings for consumer reviews. International Journal of Information Management, 54, 102132.

HYNDMAN, R. J. & ATHANASOPOULOS, G. 2018. Forecasting: principles and practice, OTexts.

PIRBAZARI, A. M., SHARMA, E., CHAKRAVORTY, A., ELMENREICH, W. & RONG, C. 2021. An Ensemble Approach for Multi-Step Ahead Energy Forecasting of Household Communities. IEEE Access, 9, 36218-36240.

SIEVERT, C. 2014. LDAvis: A method for visualising and interpreting topics.

Appendix

	1	2	3	4	V5
730	0.353448275862069	0.21551724137931	0.21551724137931	0.21551724137931	Phone is like anything . Fully satisfied with the product and not issues . I will recommend all to buy this device it is valua for money
91	0.341666666666667	0.208333333333333	0.241666666666667	0.208333333333333	Camera is awesome
564	0.336206896551724	0.21551724137931	0.232758620689655	0.21551724137931	Good product
960	0.325	0.258333333333333	0.208333333333333	0.208333333333333	Excellent
983	0.324561403508772	0.236842105263158	0.219298245614035	0.219298245614035	Worth for Money... Nice one
590	0.318965517241379	0.21551724137931	0.21551724137931	0.25	The rear camera is very good and the front camera is ok ok.
949	0.318965517241379	0.25	0.21551724137931	0.21551724137931	I didn't clicked good pic of the phone
821	0.318181818181818	0.227272727272727	0.227272727272727	0.227272727272727	Happy with this phone. Awesome battery life and camera quality.
658	0.314516129032258	0.201612903225806	0.282258064516129	0.201612903225806	Awesome performance...
10	0.307017543859649	0.219298245614035	0.236842105263158	0.236842105263158	Best part is battery and processor is good and value of money

Appendix1: Sample comments with high distribution on topic 1(P)

	1	2	3	4	V5
862	0.21551724137931	0.318965517241379	0.232758620689655	0.232758620689655	Good phone in 13k
467	0.241071428571429	0.3125	0.223214285714286	0.223214285714286	I just love this product 😍
178	0.223214285714286	0.3125	0.223214285714286	0.241071428571429	Good
128	0.214285714285714	0.30952380952381	0.277777777777778	0.198412698412698	Good
318	0.225	0.308333333333333	0.258333333333333	0.208333333333333	Awesome phone
222	0.236842105263158	0.307017543859649	0.219298245614035	0.236842105263158	Value for money.. middle level budget mobile. It's awesome . Especially I have to mention here amazon Delivery man . He is so kin...
105	0.231481481481481	0.305555555555556	0.231481481481481	0.231481481481481	I am using this from 15 days.
336	0.231481481481481	0.305555555555556	0.231481481481481	0.231481481481481	Good
495	0.231481481481481	0.305555555555556	0.231481481481481	0.231481481481481	Lovely color.....I just luv it... thanks to Amazon.......😍😍😍😍😘😘😘😘😘😍😍😍😍😍😍😍😍very happy with my redmi n...
203	0.2578125	0.3046875	0.2421875	0.1953125	Classy looking..Phone has really good Camera, battery backup is also good..
594	0.245454545454545	0.3	0.227272727272727	0.227272727272727	Look wise awesome. Fingerprint and face unlock response very fast. Sound quality good. Front camera not up to the mark in this ...
448	0.227272727272727	0.3	0.245454545454545	0.227272727272727	Got this legend delivered to me today on Occasion of Teacher's day. Wanted to gift my dad for his birthday this month. Advance ...
349	0.233870967741935	0.298387096774194	0.266129032258065	0.201612903225806	5/5 best mobile in this range
982	0.201612903225806	0.298387096774194	0.298387096774194	0.201612903225806	Phone is good in every purpose except the selfie camera quality....which i expect more superior
250	0.26271186440678	0.296610169491525	0.228813559322034	0.211864406779661	ITS FEATURES ARE GOOD BUT IT TOOK SOME TIME TO FIND WHAT IS WHERE. AS LIKE OTHER CELL PHONE OPTIONS ARE AVAILA...
515	0.211864406779661	0.296610169491525	0.26271186440678	0.228813559322034	1. Camera 4.7/5 rating
888	0.241071428571429	0.294642857142857	0.223214285714286	0.241071428571429	Good cam and very lite weight
287	0.223214285714286	0.294642857142857	0.223214285714286	0.258928571428571	Nice Product. A1 camera quality. Awosome Wide angle.

Appendix2: Sample comments with high distribution on topic 2(P)

	1	2	3	4	V5
393	0.219298245614035	0.236842105263158	0.324561403508772	0.219298245614035	excellent phone but still u can make big screen like around 7 inch if possible
215	0.236842105263158	0.219298245614035	0.324561403508772	0.219298245614035	very nice mobile.....best bettry life...nd good camera ....
554	0.254098360655738	0.221311475409836	0.319672131147541	0.204918032786885	Good
714	0.232758620689655	0.21551724137931	0.318965517241379	0.232758620689655	This is awesome creation of one plus.... All features are very good..those who's wants to buy good quality phone within good ran...
307	0.227272727272727	0.227272727272727	0.318181818181818	0.227272727272727	Nice camera quality
355	0.227272727272727	0.227272727272727	0.318181818181818	0.227272727272727	Camera, Bettary, Sound, Speed & smooth SUPERB. Value for money
987	0.192307692307692	0.238461538461538	0.315384615384615	0.253846153846154	One plus is just a brand
784	0.211864406779661	0.228813559322034	0.313559322033898	0.245762711864407	Camera quality is very good
546	0.228813559322034	0.228813559322034	0.313559322033898	0.228813559322034	Outstanding performance
790	0.228813559322034	0.228813559322034	0.313559322033898	0.228813559322034	Good product, but poor battery backup
375	0.198412698412698	0.214285714285714	0.30952380952381	0.277777777777778	The device is great .. but what about the gift box 2 ... I haven't received any information about it
538	0.241666666666667	0.241666666666667	0.308333333333333	0.208333333333333	I have been using Samsung for years now, but as expected One Plus's performance is just awesome. Camera and it's modes are s...
131	0.219298245614035	0.236842105263158	0.307017543859649	0.236842105263158	Nice
89	0.219298245614035	0.254385964912281	0.307017543859649	0.219298245614035	this product is good in 🔋 battery life

Appendix3: Sample comments with high distribution on topic 3(P)

	1	2	3	4	5	6	V7
115	0.224242424242424	0.151515151515152	0.151515151515152	0.16969696969697	0.151515151515152	0.151515151515152	Hanging problem in (redmi note 8) 3-5 time hang our mob...
248	0.222222222222222	0.138888888888889	0.205555555555556	0.138888888888889	0.138888888888889	0.155555555555556	Finger print is worst. Don't even recognise. Don't but this. ...
179	0.220238095238095	0.148809523809524	0.148809523809524	0.148809523809524	0.166666666666667	0.166666666666667	I am OnePlus fan since launch of one plus one, but I had v...
333	0.220238095238095	0.166666666666667	0.148809523809524	0.148809523809524	0.148809523809524	0.166666666666667	Battery drains as fast as it charges. It fully charges in 30 mi...
198	0.218579234972678	0.169398907103825	0.153005464480874	0.136612021857924	0.153005464480874	0.169398907103825	Green Tint And Pink Tint Problem
88	0.216374269005848	0.146198830409357	0.16374269005848	0.146198830409357	0.16374269005848	0.16374269005848	Not bad
239	0.216374269005848	0.146198830409357	0.146198830409357	0.146198830409357	0.16374269005848	0.181286549707602	Not a product
299	0.216374269005848	0.146198830409357	0.146198830409357	0.16374269005848	0.146198830409357	0.181286549707602	Camera too worst... & Slow motion camera was dim light ...
398	0.216374269005848	0.16374269005848	0.146198830409357	0.146198830409357	0.146198830409357	0.181286549707602	Worst thing to buy from China waste of money
22	0.213836477987421	0.157232704402516	0.157232704402516	0.157232704402516	0.157232704402516	0.157232704402516	Given 64MP Camra quality is too bad. Battery is drening fa...
149	0.213836477987421	0.157232704402516	0.157232704402516	0.157232704402516	0.157232704402516	0.157232704402516	Average
49	0.21264367816092	0.14367816091954	0.195402298850575	0.14367816091954	0.14367816091954	0.160919540229885	1+ loosing their standard.
131	0.21264367816092	0.14367816091954	0.17816091954023	0.17816091954023	0.14367816091954	0.14367816091954	Instagram not working on this device
27	0.209876543209877	0.154320987654321	0.154320987654321	0.154320987654321	0.154320987654321	0.17283950617284	A
175	0.209876543209877	0.154320987654321	0.154320987654321	0.17283950617284	0.154320987654321	0.154320987654321	The phone keeps lagging and gets off itself.IIII Very bad e...
200	0.209876543209877	0.154320987654321	0.154320987654321	0.154320987654321	0.154320987654321	0.17283950617284	I faced network signal issue. I don't know what to do.
509	0.209876543209877	0.154320987654321	0.154320987654321	0.17283950617284	0.154320987654321	0.154320987654321	Below average image quality. Selfies come with a halo aro...

Appendix4: Sample comments with high distribution on topic 1(N)

	1	2	3	4	5	6	V7
196	0.154320987654321	0.228395061728395	0.154320987654321	0.154320987654321	0.154320987654321	0.154320987654321	One of my biggest bad decision to buy oneplus nord. One o...
282	0.154320987654321	0.228395061728395	0.154320987654321	0.154320987654321	0.154320987654321	0.154320987654321	Very poor battery life
8	0.151515151515152	0.224242424242424	0.16969696969697	0.151515151515152	0.151515151515152	0.151515151515152	don't buy this phone. waste of money
325	0.16969696969697	0.224242424242424	0.151515151515152	0.151515151515152	0.151515151515152	0.151515151515152	The phone is simple and cool
96	0.151515151515152	0.224242424242424	0.151515151515152	0.16969696969697	0.151515151515152	0.151515151515152	Ones you Full charged for a day good enough,which I receiv...
48	0.185792349726776	0.218579234972678	0.136612021857924	0.136612021857924	0.153005464480874	0.169398907103825	It does not support mobile data for video and PUBG . The c...
231	0.153005464480874	0.218579234972678	0.136612021857924	0.202185792349727	0.136612021857924	0.153005464480874	1. Battery usage update: Drains faster than other one plus m...
273	0.16374269005848	0.216374269005848	0.146198830409357	0.16374269005848	0.146198830409357	0.16374269005848	very bed camera performence also battery backup
227	0.146198830409357	0.216374269005848	0.146198830409357	0.181286549707602	0.16374269005848	0.146198830409357	After one month intense use I am here to submit feedback.....
497	0.157232704402516	0.213836477987421	0.157232704402516	0.157232704402516	0.157232704402516	0.157232704402516	Not so good not so bad and phone is very weighted
516	0.157232704402516	0.213836477987421	0.157232704402516	0.157232704402516	0.157232704402516	0.157232704402516	I don't like this phone
441	0.14367816091954	0.21264367816092	0.17816091954023	0.14367816091954	0.14367816091954	0.17816091954023	I am using dual sims. after every 2-3 hours, outgoing and in...
424	0.14367816091954	0.21264367816092	0.160919540229885	0.160919540229885	0.160919540229885	0.160919540229885	BAD BATTERY LIFE OR MOBILE HANG PROBLEM
59	0.154320987654321	0.209876543209877	0.17283950617284	0.154320987654321	0.154320987654321	0.154320987654321	Ok
89	0.154320987654321	0.209876543209877	0.17283950617284	0.154320987654321	0.154320987654321	0.154320987654321	Hang
352	0.141242937853107	0.209039548022599	0.141242937853107	0.141242937853107	0.175141242937853	0.192090395480226	Battery backup is not good. My full charge battery only wor...
379	0.151515151515152	0.206060606060606	0.16969696969697	0.151515151515152	0.16969696969697	0.151515151515152	All the hype created got shattered in seconds. Not at all wor...

Appendix5: Sample comments with high distribution on topic 2(N)

	1	2	3	4	5	6	V7
195	0.166666666666667	0.148809523809524	0.220238095238095	0.148809523809524	0.166666666666667	0.148809523809524	Phone is getting heat while charging. Redmi has to look in t...
265	0.160919540229885	0.14367816091954	0.21264367816092	0.160919540229885	0.17816091954023	0.14367816091954	Phone is good , but battery very bad and no battery life 😞 ...
127	0.14367816091954	0.14367816091954	0.21264367816092	0.160919540229885	0.17816091954023	0.160919540229885	Received the damaged product highly dissatisfied
26	0.154320987654321	0.17283950617284	0.209876543209877	0.154320987654321	0.154320987654321	0.154320987654321	Battery life is excellent but camera bad and fingerprint sens...
110	0.154320987654321	0.154320987654321	0.209876543209877	0.154320987654321	0.17283950617284	0.154320987654321	📷 quality not good. Performance Very bad. Worst of money.
100	0.151515151515152	0.187878787878788	0.206060606060606	0.151515151515152	0.151515151515152	0.151515151515152	Defective product. Mic not working and battery getting char...
492	0.151515151515152	0.16969696969697	0.206060606060606	0.151515151515152	0.16969696969697	0.151515151515152	Bad
406	0.16969696969697	0.151515151515152	0.206060606060606	0.16969696969697	0.151515151515152	0.151515151515152	Good
185	0.151515151515152	0.151515151515152	0.206060606060606	0.151515151515152	0.187878787878788	0.151515151515152	Battery draining very fast,even an aeroplane mode battery d...
238	0.151515151515152	0.151515151515152	0.206060606060606	0.16969696969697	0.151515151515152	0.16969696969697	The worst purchase ever made. If you are looking for a came...
248	0.222222222222222	0.138888888888889	0.205555555555556	0.138888888888889	0.138888888888889	0.155555555555556	Finger print is worst. Don't even recognise. Don't but this. Ev...
270	0.18452380952381	0.166666666666667	0.202380952380952	0.148809523809524	0.148809523809524	0.148809523809524	We
97	0.166666666666667	0.148809523809524	0.202380952380952	0.148809523809524	0.18452380952381	0.148809523809524	Total waste of money don't purchase
126	0.148809523809524	0.148809523809524	0.202380952380952	0.148809523809524	0.202380952380952	0.148809523809524	Proximity sensor is not working anymore
439	0.148809523809524	0.148809523809524	0.202380952380952	0.148809523809524	0.202380952380952	0.148809523809524	Call system is so bad . Call us automatically cut in 5-10 sec . ...

Appendix6: Sample comments with high distribution on topic 3(N)