

Assignment - 1

The given dataset consists of 2358 rows with 6 columns. The 'id' column specifies the unique id of each row, but when the dataset is loaded through pandas it automatically assigns a unique id, so the existing id column is dropped. The 'from' and 'to' columns reveal the position of the aspect term in the sentence, but the position of the aspect term can change during preprocessing, so I dropped the 'from' and 'to' columns from the dataset.

	Sentence	Aspect Term	polarity
0	I charge it at night and skip taking the cord ...	cord	neutral
1	I charge it at night and skip taking the cord ...	battery life	positive
2	The tech guy then said the service center does...	service center	negative
3	The tech guy then said the service center does...	"sales" team	negative
4	The tech guy then said the service center does...	tech guy	neutral

Fig. 1 First five rows after dropping three columns.

The 'polarity' column has four different unique values namely positive, negative, neutral, and conflict.

```
polarity
positive    987
negative    866
neutral     460
conflict     45
Name: count, dtype: int64
```

Figure. 2 Different class values of the 'polarity'

1. Lowercasing

In the first part of preprocessing the sentences were lowercased along with the aspect term as aspect term is crucial for determining the polarity of the sentence.

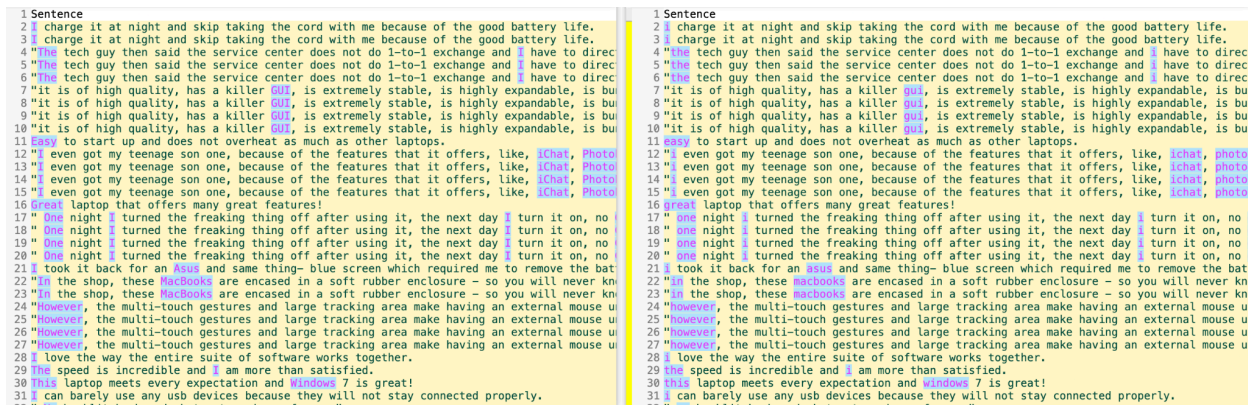


Figure 3. Before and After lowercasing sentences

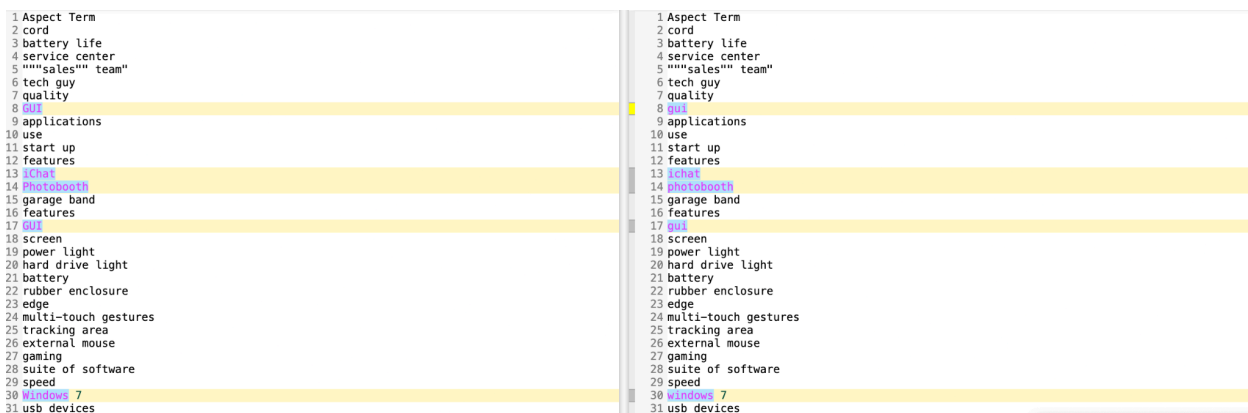


Figure 4. Before and After lowercasing aspect term

2. Removal of URLs, Email Address, and Hashtags

In the second step, sentences were inspected for URLs, Email Address, and Hashtags, but it lacks any of these in it.

3. Removal of white space

This is a common preprocessing task in NLP. In this step we removed extra space and leading/trailing spaces in the sentences and aspect term.

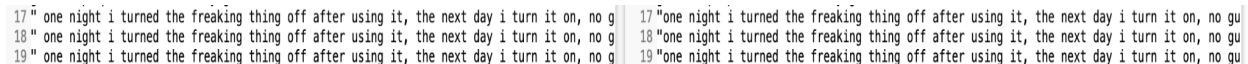


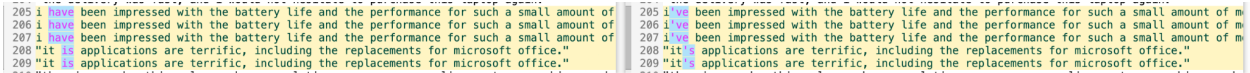
Figure 5. Before and After removal of white space in sentences



Figure 6. Before and After removal of white space in aspect term

4. Handling Contractions

Here we expand the contractions to their full forms. It helps to standardize the text and can improve the performance.

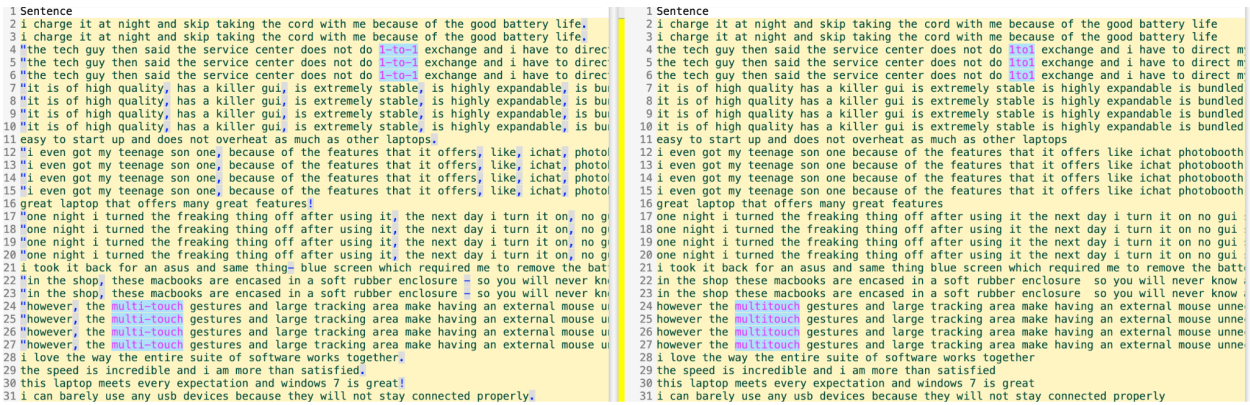


205 i have been impressed with the battery life and the performance for such a small amount of m
206 i have been impressed with the battery life and the performance for such a small amount of m
207 i have been impressed with the battery life and the performance for such a small amount of m
208 it is applications are terrific, including the replacements for microsoft office."
209 it is applications are terrific, including the replacements for microsoft office."

Figure 7. After and Before expanding the contractions.


5. Removal of special characters.

Special characters are removed to reduce noise, and prevent unnecessary tokens from inflating the vocabulary size.



1 Sentence
2 i charge it at night and skip taking the cord with me because of the good battery life.
3 i charge it at night and skip taking the cord with me because of the good battery life.
4 the tech guy then said the service center does not do i-to-i exchange and i have to direc
5 the tech guy then said the service center does not do i-to-i exchange and i have to direc
6 the tech guy then said the service center does not do i-to-i exchange and i have to direc
7 it is of high quality, has a killer gui, is extremely stable, is highly expandable, is bu
8 it is of high quality, has a killer gui, is extremely stable, is highly expandable, is bu
9 it is of high quality, has a killer gui, is extremely stable, is highly expandable, is bu
10 it is of high quality, has a killer gui, is extremely stable, is highly expandable, is bu
11 easy to start up and does not overheat as much as other laptops.
12 i even got my teenage son one, because of the features that it offers, like, ichat, photol
13 i even got my teenage son one, because of the features that it offers, like, ichat, photol
14 i even got my teenage son one, because of the features that it offers, like, ichat, photol
15 i even got my teenage son one, because of the features that it offers, like, ichat, photol
16 great laptop that offers many great features!
17 one night i turned the freaking thing off after using it, the next day i turn it on, no g
18 one night i turned the freaking thing off after using it, the next day i turn it on, no g
19 one night i turned the freaking thing off after using it, the next day i turn it on, no g
20 one night i turned the freaking thing off after using it, the next day i turn it on, no g
21 i took it back for an asus and same thing blue screen which required me to remove the batt
22 in the shop, these macbooks are encased in a soft rubber enclosure = so you will never kn
23 in the shop, these macbooks are encased in a soft rubber enclosure = so you will never kn
24 however, the multi-touch gestures and large tracking area make having an external mouse u
25 however, the multi-touch gestures and large tracking area make having an external mouse u
26 however, the multi-touch gestures and large tracking area make having an external mouse u
27 however, the multi-touch gestures and large tracking area make having an external mouse u
28 i love the way the entire suite of software works together.
29 the speed is incredible and i am more than satisfied.
30 this laptop meets every expectation and windows 7 is great!
31 i can barely use any usb devices because they will not stay connected properly.

Figure 8. Before and After removal of special characters in sentences.



1 Aspect Term
2 cord
3 battery life
4 service center
5 ""sales"" team"
6 tech guy
7 quality
8 gui
9 applications
10 use
11 start up
12 features
13 ichat
14 photobooth
15 garage band
16 features
17 gui
18 screen
19 power light
20 hard drive light
21 battery
22 rubber enclosure
23 edge
24 multi-touch gestures
25 tracking area
26 external mouse
27 gaming
28 suite of software
29 speed

Figure 9. Before and After removal of special characters in aspect term.

6. Handling numbers.

Numbers are removed to reduce noise, improve generalization, and simplify the text data.

52 in desperation i called their customer service number and was told that my warranty had ex
53 in desperation i called their customer service number and was told that my warranty had ex
54 in desperation i called their customer service number and was told that my warranty had ex
55 there also seemed to be a problem with the hard disc as certain times windows loads but cl
56 there also seemed to be a problem with the hard disc as certain times windows loads but cl
57 there also seemed to be a problem with the hard disc as certain times windows loads but cl
58 drivers updated ok but the bios update froze the system up and the computer shut down
59 drivers updated ok but the bios update froze the system up and the computer shut down
60 drivers updated ok but the bios update froze the system up and the computer shut down
61 spent hours on phone with hp technical support
62 speaking of the browser it too has problems
63 the keyboard is too slick
64 nightly my computer defrags itself and runs a virus scan
65 it is like punds but if you can look past it it is great
66 it is just as fast with one program open as it is with sixteen open
67 still under warrenty so called toshiba no help at all
68 i was happy with my purchase of a toshiba satellite 145 laptop until it came time to have
69 amazing quality
70 the fact that you can spend over on just a webcam underscores the value of this machine
71 the fact that you can spend over on just a webcam underscores the value of this machine
72 i mainly use it for email internet and managing personal files pics vids etc
73 i mainly use it for email internet and managing personal files pics vids etc
74 a month or so ago the freaking motherboard just died
75 the system it comes with does not work properly so when trying to fix the problems with it
76 then after or so months the charger stopped working so i was forced to go out and buy new
77 then after or so months the charger stopped working so i was forced to go out and buy new
78 if a website ever freezes which is rare its really easy to force quit
79 it rarely works and when it does it is incredibly slow
80 i also enjoy the fact that my macbook pro laptop allows me to run windows on it by using
81 i also enjoy the fact that my macbook pro laptop allows me to run windows on it by using
82 it is so much easier to navigate through the operating system to find files and it runs a

Figure 10. After and Before handling numbers in sentences.

30 windows
31 usb devices
32 keyboard
33 software
34 system
35 microsoft office for the mac
36 syncing
37 hd monitor

Figure 11. After and Before handling numbers in aspect term.

7. Tokenization

It involves breaking down text into smaller units, known as tokens.

'i charge it at night and skip taking the cord with me because of the good battery life'

Figure 12. Sentence Before Tokenization

```
['i',  
'charge',  
'it',  
'at',  
'night',  
'and',  
'skip',  
'taking',  
'the',  
'cord',  
'with',  
'me',  
'because',  
'of',  
'the',  
'good',  
'battery',  
'life']
```

Figure 13. Sentence After Tokenization

8. Stop Word removal

Stop words are common words such as "the," "is," "in," "and," and "to," which occur frequently in text but carry little meaning. This step helps to reduce noise and improves model efficiency.

```
['charge', 'night', 'skip', 'taking', 'cord', 'good', 'battery', 'life']
```

Figure 14. Result After Stop Word Removal

9. Lemmatization

The purpose of lemmatization is to group different forms of a word into a single base form so that they can be treated as the same word. This was chosen over stemming as it results in better accuracy.

```
['charge', 'night', 'skip', 'take', 'cord', 'good', 'battery', 'life']
```

Figure 14. Result After Lemmatization

```
0      charge night skip take cord good battery life
1      charge night skip take cord good battery life
2  tech guy say service center exchange direct co...
3  tech guy say service center exchange direct co...
4  tech guy say service center exchange direct co...
..
```

Figure 15. Dataset after preprocessing.

Bag of Words:

This is the first frequency based method that is used on the dataset. Before applying this method on the dataset, the aspect term column is combined with the sentence column as this approach is more straightforward and can work well when the aspect term naturally fits within the context of the sentence. Even though using separate embeddings and concatenating them gives more flexibility, it results in poor accuracy.

CountVectorizer from sklearn is used for Bag of Words and chi2 is used to select top 100 features. The result from the chi-square is used to split the dataset into train and test data. Before train-test split of the data the polarity column of the dataset is label encoded.

Three classifiers Random Forest, Decision Tree, SVC used on the embedding. The result can be seen in the table below.

Classifier	Accuracy
Random Forest Classifier	61.86%
SVC	62.50%
Decision Tree Classifier	61.86%

The following results are obtained during cross validation for this embedding.

Classifier	Cross Validation Accuracy
Random Forest Classifier	59.54%
SVC	60.43%
Decision Tree Classifier	58.53%

TF - IDF

This is the second frequency based method that is used on the dataset. Before applying this method on the dataset, the aspect term column is combined with the sentence column as this approach is more straightforward and can work well when the aspect term naturally fits within the context of the sentence. Even though using separate embeddings and concatenating them gives more flexibility, it results in poor accuracy.

TfidfVectorizer from sklearn is used for TF-IDF and chi2 is used to select top 100 features. The result from the chi-square is used to split the dataset into train and test data. Before train-test split of the data the polarity column of the dataset is label encoded.

Three classifiers Random Forest, Decision Tree, SVC used on the embedding. The result can be seen in the table below.

Classifier	Accuracy
Random Forest Classifier	60.80%
SVC	61.44%
Decision Tree Classifier	60.16%

The following results are obtained during cross validation for this embedding.

Classifier	Cross Validation Accuracy
Random Forest Classifier	56.83%
SVC	59.46%
Decision Tree Classifier	55.64%

For the frequency based approach Bag of Words dominates the TF-IDF method as the accuracy and cross validation accuracy for the former is higher than latter.

GloVe

For this embedding method glove.6B.300d is used from stanford. As a part of this embedding process both sentence and aspect column of the dataset is passed separately to the method to capture embedding for the words in sentence and aspect column. After this the result from the sentence and aspect embedding were combined to form a final embedding. The final embedding is used to obtain train and test data.

Three classifiers Random Forest, Decision Tree, SVC used on the embedding. The result can be seen in the table below.

Classifier	Accuracy
Random Forest Classifier	68.64%
SVC	56.99%
Decision Tree Classifier	54.66%

The following results are obtained during cross validation for this embedding.

Classifier	Cross Validation Accuracy
Random Forest Classifier	62.09%
SVC	60.05%
Decision Tree Classifier	48.14%

Random Forest's superior performance with GloVe embeddings can be attributed to its ability to handle high-dimensional and complex data. Unlike SVC, which may struggle with non-linearly separable data, and Decision Trees, which tend to overfit, Random Forest reduces overfitting by averaging multiple trees.

Sentence BERT

For this embedding all-distilroberta-v1 is used from sentence-transformers. At first embedding is applied separately on the sentence and aspect column of the dataset. The result is concatenated to form a combined embedding. The final embedding is used to obtain train and test data. Three classifiers Random Forest, Decision Tree, SVC used on the embedding. The result can be seen in the table below.

Classifier	Accuracy
Random Forest Classifier	69.49%
SVC	67.50%
Decision Tree Classifier	57.74%

The following results are obtained during cross validation for this embedding.

Classifier	Cross Validation Accuracy
Random Forest Classifier	64.25%
SVC	66.80%
Decision Tree Classifier	49.37%

The better performance of BERT over GloVe and Bag of Words can be attributed to the differences in how these models capture the meaning of words in context. BERT is a contextualized word embedding model, meaning it understands the meaning of words based on their surrounding context within a sentence. In contrast, GloVe is a static word embedding model, meaning that each word has a fixed vector representation regardless of context, which limits its ability to capture nuanced meanings in different sentences. Bag of Words is even more simplistic, treating words independently without considering their context or relationships, leading to a less expressive representation of the text.

References:

<https://nlp.stanford.edu/projects/glove/>

<https://medium.com/analytics-vidhya/basics-of-using-pre-trained-glove-vectors-in-python-d38905f356db>

<https://www.geeksforgeeks.org/python-lemmatization-with-nltk/>

<https://medium.com/analytics-vidhya/the-best-feature-selection-technique-for-text-classification-23199b4a4f8d>

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html

<https://huggingface.co/sentence-transformers/all-mpnet-base-v2>