

JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY



Title : Sentiment Analysis of Reviews using KNN

B. Tech - SEMESTER-V

Fundamentals of Machine Learning

SUBMITTED TO: Dr. Dhanalakshmi

<u>GROUP MEMBERS :</u>	<u>BATCH</u>
-------------------------------	---------------------

ADIT SHARMA: 20103301	B1
-----------------------	----

NISHTHA GULATI: 20103100	B4
--------------------------	----

HARSHIT CHOPRA: 20104013	B14
--------------------------	-----

AYUSH SHARMA: 20104059	B14
------------------------	-----

TABLE OF CONTENTS

Problem statement	2
Algorithm used	2
Motivation	2
Literature available on the problem	3
Model used/proposed (flow diagram)	4 - 5
Implementation	6 - 10
Results	11 - 13
Conclusion	13
References	13

PROBLEM STATEMENT

A practical application in e-commerce applications is to infer sentiment (or polarity) from free-form review text submitted for a range of products. We will implement the k-Nearest Neighbor algorithm to predict the sentiment for the reviews for various products.

ALGORITHM USED

The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand.

KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression).

MOTIVATION

Social networking platforms have been gaining huge popularity amongst people in recent years. These online sites are being widely used by people to express their emotions, beliefs as well as opinions towards any entity ranging from product, person, event and so on. These networking sites provide a platform for users to post their feedback and reviews and the data generated therein are being harnessed by business enterprises in order to get an insight into how well their products and services are faring in the market.

This knowledge helps business analysts and managers in better decision making. Apart from business enterprises, sentiment analysis of user comments is of immense use for buyers too.

However it is not possible for a user to analyze all the reviews considering the massive amount of user reviews and comments available on online platforms. Hence our sentiment analysis technique has been proposed in order to automate this analysis process. Through this technique a user would be able to know about the positive as well as negative views that the other users have regarding a product. Thus the user gets a clear view about the products and services, and can assess whether it is as per the requirements.

LITERATURE AVAILABLE ON THE PROBLEM

Sumbal Riaz et al. recommended an approach termed text mining for examining customer reviews to ascertain the customers' opinions and executed the SA on the massive dataset of product (6 sorts) reviews proffered by disparate customers on the internet. In this approach, SA was employed at the phrase level instead of document-level for computing every term's SP. Then key graph keyword extraction was used aimed at extracting keywords as of each document with high-frequency terms and the intensity of SP by gauging its strength was evaluated. The k-means clustering was utilized for grouping data on the base of sentiment strength value. Those values were contrasted to the star rating of the same data and the excellent and neutral sentiment toward products was examined. The approach uses clustering which may bring about over clustering.

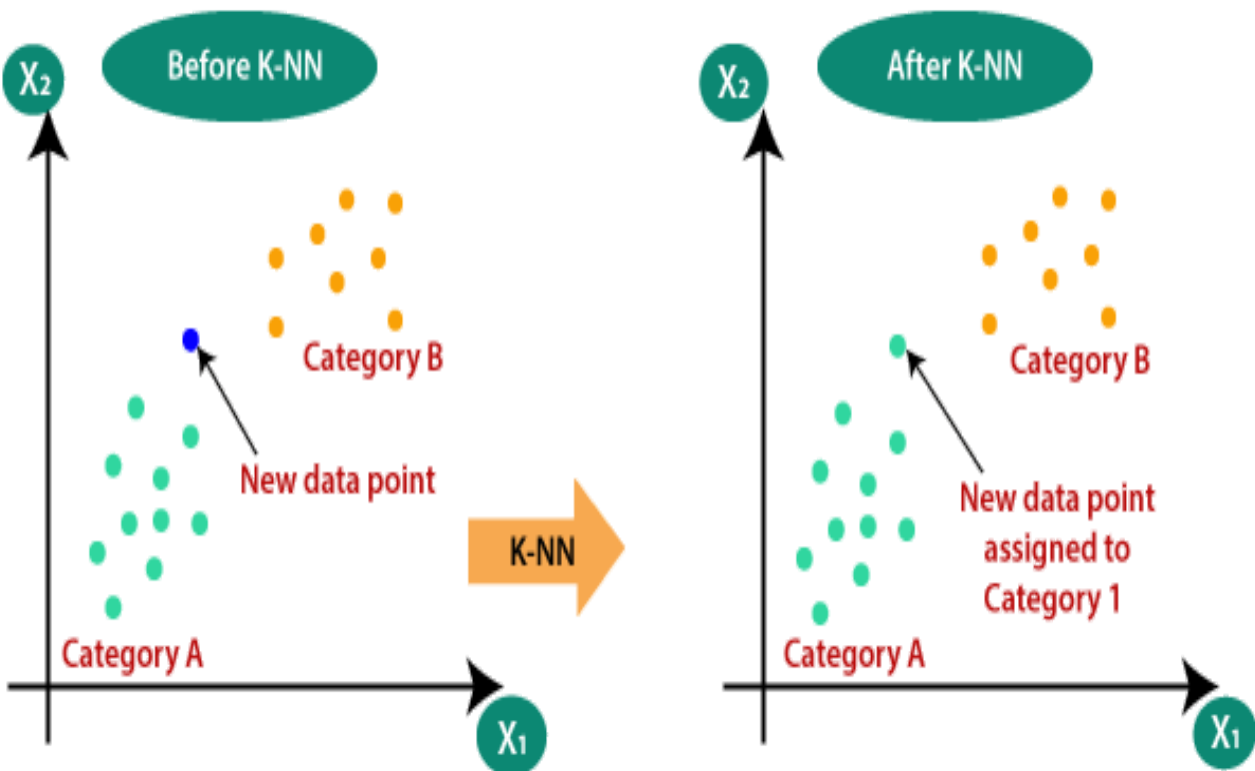
Feilong Tang et al. suggested the '2' generative model, MaxEnt-JABST as well as JABST, that extracted typically the fine-grained opinions along with aspects as of reviews (online). The JABST model extracted particular and general opinions and aspects together with the sentiment polarity (SP). In addition, the MaxEnt-JABST design added a maximal entropy classifier for separating aspects or opinion words more precisely. Those designs were assessed on review regarding restaurants and electronic devices quantitatively as well as qualitatively. The experiential outcomes evinced that the designs outperformed existent baselines and were competent to recognize fine-grained aspects and opinions but the improvement was still needed.

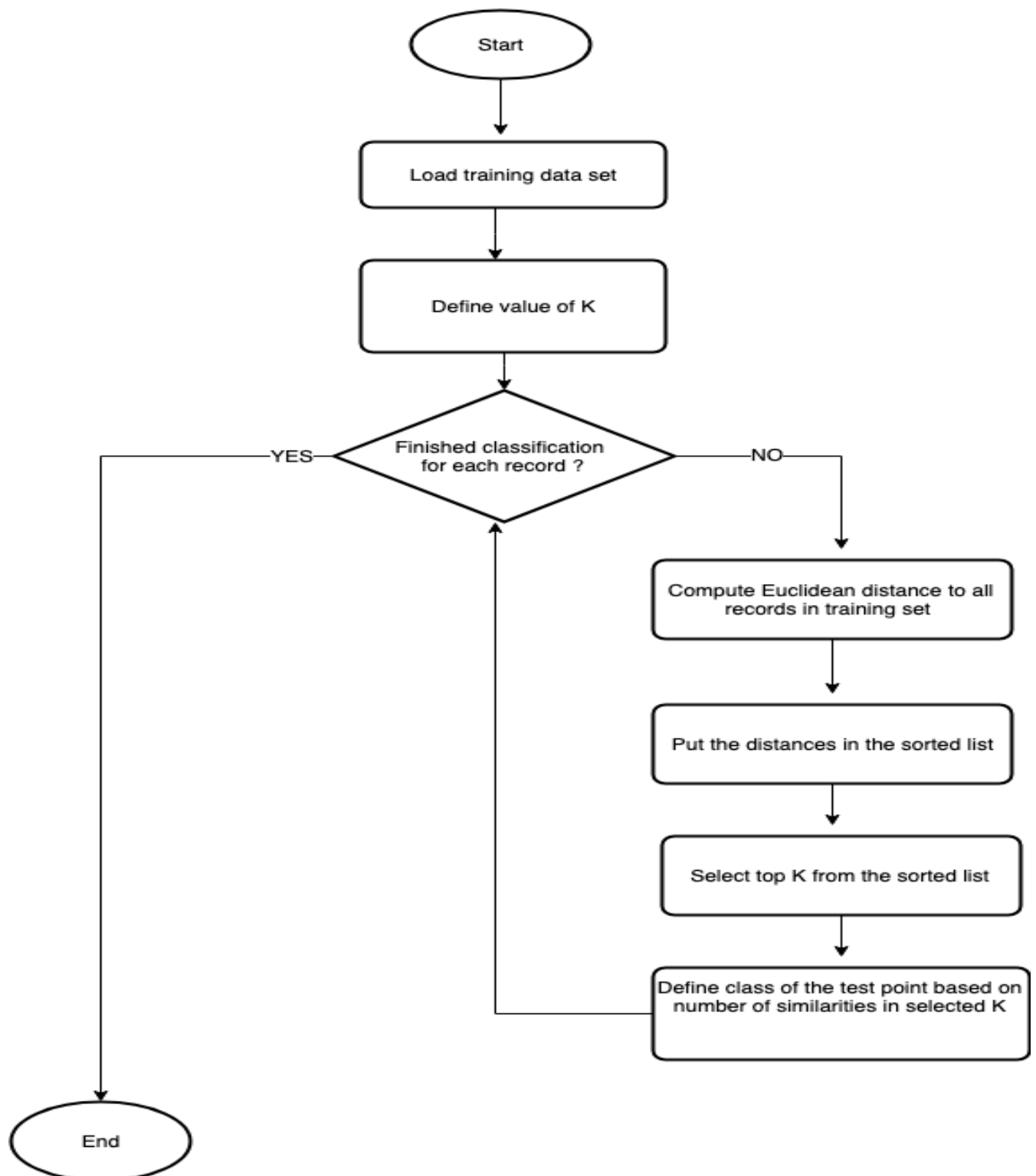
Rajkumar et al. rendered a '2' ML approach, say Naïve Bayes (NB) and SVM for performing SA on reviews of a specific product. In those approaches, the dataset was gathered as of Amazon, which comprised reviews regarding Laptops, Cameras, Mobiles, Tablets, video surveillance, and TVs. Subsequently, stemming, stop word removal, and also punctuation marks removal were executed and it was transmuted into a bag of words. This dataset was contrasted to opinion lexicons, that is, 4783 negative and 2006 positive words with sentiment scores intended for every sentence were evaluated. Utilizing score and disparate features, the NB along with SVM were employed and diverse accurateness was computed

MODEL USED/PROPOSED (FLOW DIAGRAM)

The k-nearest neighbors algorithm, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression).





Implementation

1. Initially we created our own k-nearest neighbors class and tested its accuracy using iris data set.

MODEL CREATION

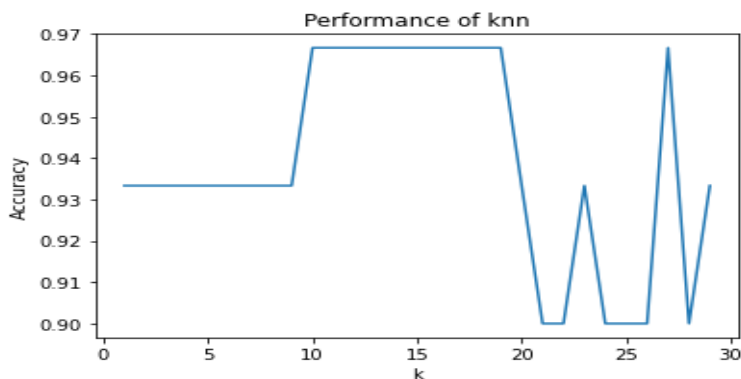
```
class KneighborsClassifier:
    def __init__(self, k=5, dist_metric=euclidean):
        self.k = k
        self.dist_metric = dist_metric

    def fit(self, X_train, y_train):
        self.X_train = X_train
        self.y_train = y_train

    def predict(self, X_test):
        neighbors = []
        for x in X_test:
            distances = self.dist_metric(x, self.X_train)
            y_sorted = [y for _, y in sorted(zip(distances, self.y_train))]
            neighbors.append(y_sorted[:self.k])
        return list(map(most_common, neighbors))

    def evaluate(self, X_test, y_test):
        y_pred = self.predict(X_test)
        accuracy = sum(y_pred == y_test) / len(y_test)
        return accuracy
```

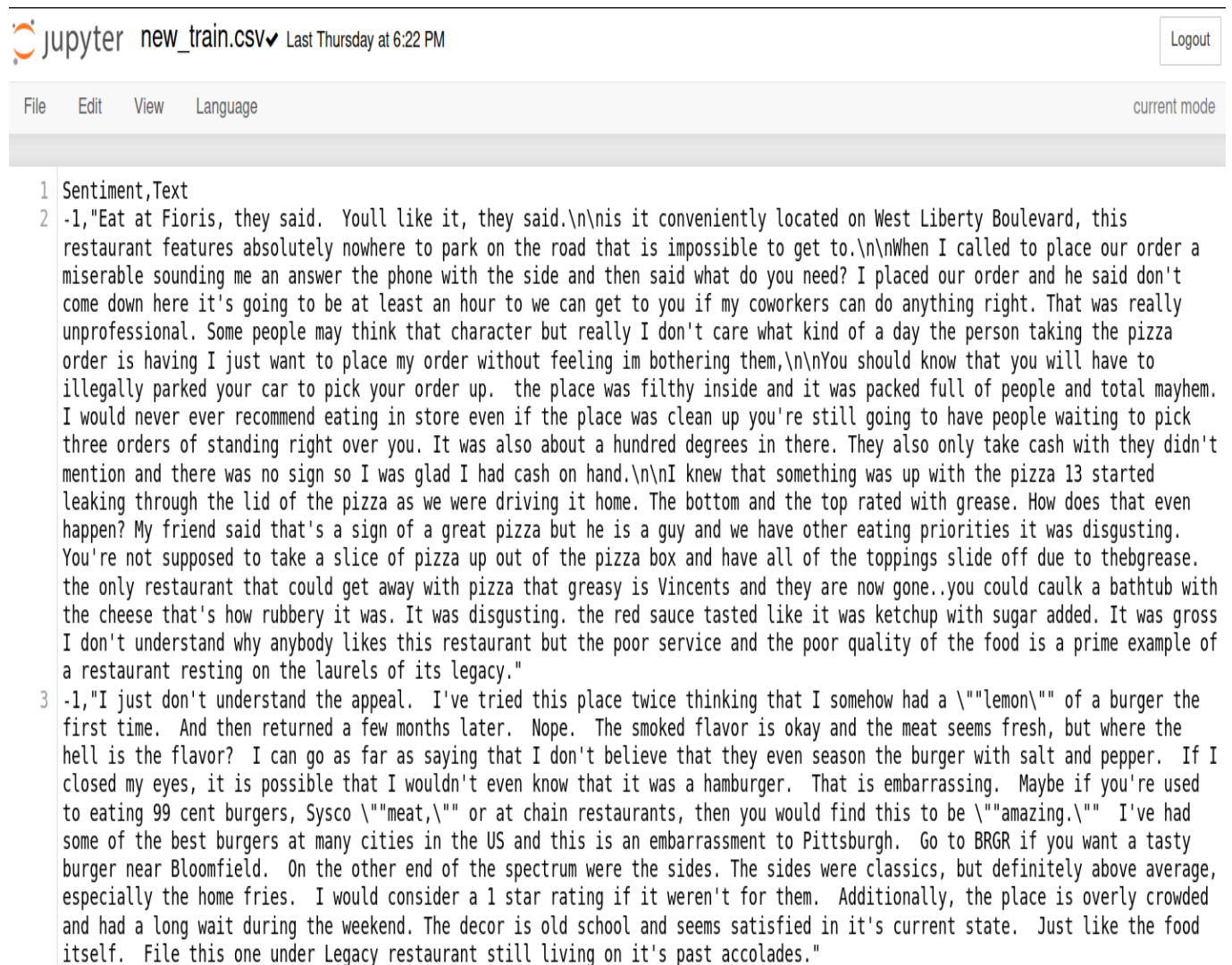
The data is split into training and test sets. It is then preprocessed and the accuracy measured.



2. Now the training and testing is performed on actual reviews stored in new_train.csv and new_test.csv

Training Dataset

1000 reviews exist in the file new_train.csv. Each row begins with the sentiment score followed by the text of the review.



```
1 Sentiment,Text
2 -1,"Eat at Fioris, they said. Youll like it, they said.\n\nis it conveniently located on West Liberty Boulevard, this
restaurant features absolutely nowhere to park on the road that is impossible to get to.\n\nWhen I called to place our order a
miserable sounding me an answer the phone with the side and then said what do you need? I placed our order and he said don't
come down here it's going to be at least an hour to we can get to you if my coworkers can do anything right. That was really
unprofessional. Some people may think that character but really I don't care what kind of a day the person taking the pizza
order is having I just want to place my order without feeling im bothering them,\n\nYou should know that you will have to
illegally parked your car to pick your order up. the place was filthy inside and it was packed full of people and total mayhem.
I would never ever recommend eating in store even if the place was clean up you're still going to have people waiting to pick
three orders of standing right over you. It was also about a hundred degrees in there. They also only take cash with they didn't
mention and there was no sign so I was glad I had cash on hand.\n\nI knew that something was up with the pizza 13 started
leaking through the lid of the pizza as we were driving it home. The bottom and the top rated with grease. How does that even
happen? My friend said that's a sign of a great pizza but he is a guy and we have other eating priorities it was disgusting.
You're not supposed to take a slice of pizza up out of the pizza box and have all of the toppings slide off due to thebgrease.
the only restaurant that could get away with pizza that greasy is Vincents and they are now gone..you could caulk a bathtub with
the cheese that's how rubbery it was. It was disgusting. the red sauce tasted like it was ketchup with sugar added. It was gross
I don't understand why anybody likes this restaurant but the poor service and the poor quality of the food is a prime example of
a restaurant resting on the laurels of its legacy."
3 -1,"I just don't understand the appeal. I've tried this place twice thinking that I somehow had a \'"lemon\'" of a burger the
first time. And then returned a few months later. Nope. The smoked flavor is okay and the meat seems fresh, but where the
hell is the flavor? I can go as far as saying that I don't believe that they even season the burger with salt and pepper. If I
closed my eyes, it is possible that I wouldn't even know that it was a hamburger. That is embarrassing. Maybe if you're used
to eating 99 cent burgers, Sysco \'"meat,\'" or at chain restaurants, then you would find this to be \'"amazing.\'" I've had
some of the best burgers at many cities in the US and this is an embarrassment to Pittsburgh. Go to BRGR if you want a tasty
burger near Bloomfield. On the other end of the spectrum were the sides. The sides were classics, but definitely above average,
especially the home fries. I would consider a 1 star rating if it weren't for them. Additionally, the place is overly crowded
and had a long wait during the weekend. The decor is old school and seems satisfied in it's current state. Just like the food
itself. File this one under Legacy restaurant still living on it's past accolades."
```


Testing Dataset

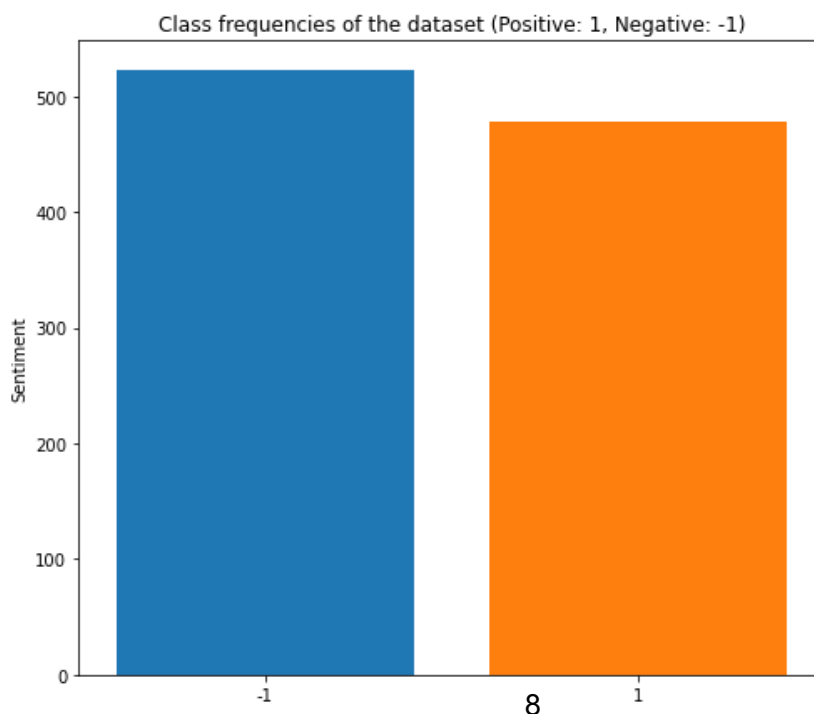
1000 reviews exist in the file new_test.csv. Each row contains the text of the review.

```
jupyter new_test.csv ✓ an hour ago Logout
```

```
File Edit View Language current mode
```

```
1 Text
2 Got take-out from here last night and it was HORRIBLE! Something must have happened because they only had 3 people working and
  were VERY disorganized. Should have checked the food before we left! Ordered fajitas and there was very little meat... it was
  beyond ridiculous. Hardly any peppers but a plethora of what looked like stewed tomatoes. Look... I'm a gringa and I know that
  don't look right and I could do a way better job. The rice was missing completely and there was only one giant pizza pie looking
  tortilla. We called and the person who answered could barely speak English (which is really not a big deal in my book) but there
  was no manager around to resolve the situation. We asked for an email so we could send a note about the experience but she was
  unhelpful. All she offered was for us to come back right then...which if you were as hungry as my fiancé was not an option. He
  gets grumpy and I know better than mess with a grumpy-hungry-pants.
3 Girls are sweet and prices are reasonable. The stand up bed is hot so make sure you adjust your time ( I might just be super
  white lol)
4 Rudest people I have ever encountered. Husband and wife owned business and when I called for service the wife was unbelievably
  and unnecessarily rude and demanding. Don't waste your time calling.
5 "This airport is only coveted for the destination that it leads its fliers too... the view of the Strip is the only thing that
  redeems this hotel. The tacky slot machines in the center of the terminal is one thing, but they don't even have a smoker's
  lounge (its vegas for god's sake!) There are not that many options for food at the airport. Also, it is a TINY airport with a
  lot of traffic and THE TRAFFIC LINE OF DOOM in itself is a catastrophe. To its defense, you can drink while you are in line,
  which i saw many people doing."
```

Class frequencies of the new_train.csv dataset



PRE-PROCESSING

3. Text-Processing

- a. Remove symbols(';', '-', ...etc)
- b. Remove stop words
- c. Stemming

```
import nltk
nltk.download('stopwords')
nltk.download('punkt')
```

```
import re, string, unicodedata
from string import punctuation
from nltk.corpus import stopwords
```

```
stop_words = set(stopwords.words("english"))
punctuation = list(string.punctuation)
stop_words.update(punctuation)
```

```
from bs4 import BeautifulSoup
def cleanText(text):
    text = BeautifulSoup(text, "lxml").text
    text = re.sub(r'\\|\\|\\|', r' ', text)
    text = re.sub(r'http\S+', r'<URL>', text) #removing urls
    text = text.lower()
    text = text.replace('x', '')
    return text
train['Text'] = train['Text'].apply(cleanText)
```

```
train['Text'] = train['Text'].str.replace('[^\w\s]', '') #removing punctuation marks
train['Text'] = train['Text'].str.replace('\d', '') #removing numbers
train['Text'] = train['Text'].apply(lambda x: " ".join(x for x in x.split() if x not in stop_words)) #stopwords removed
```

4. Splitting data into training and testing sets (test size = 0.3)

Splitting Data

```
X=train['Text']  
Y=train['Sentiment']  
  
X_train,X_test,y_train,y_test=train_test_split(X,Y,test_size=0.30,random_state=0)
```

5. Vectorization

To get some distinct features out of the text for the model to train on, by converting text to numerical vectors.

```
vectorization = TfidfVectorizer()  
Xv_train = vectorization.fit_transform(X_train)  
Xv_test = vectorization.transform(X_test)
```

6. Performing KNN Classification (n_neighbors = 20)

Training and testing

```
from sklearn.neighbors import KNeighborsClassifier  
knn=KNeighborsClassifier(n_neighbors=20,)
```

```
knn.fit(Xv_train,y_train)
```

```
KNeighborsClassifier(n_neighbors=20)
```

\

Testing Model

```
result=[]  
X_test = X_test.apply(cleanText)  
x_test = vectorization.transform(X_test)  
result= knn.predict(x_test)
```

RESULTS

Predicted Result after testing (new_train.csv)

```
result[:100]
```

```
array([ 1, -1, -1,  1,  1,  1,  1, -1,  1, -1,  1, -1,  1,  1,  1, -1, -1,
       -1,  1, -1, -1,  1, -1, -1, -1,  1,  1,  1, -1, -1,  1,  1, -1, -1,
       -1,  1, -1, -1, -1, -1,  1, -1, -1,  1, -1, -1,  1, -1, -1, -1, -1,
       -1,  1,  1,  1, -1, -1,  1,  1,  1, -1, -1,  1, -1, -1, -1,  1, -1,
        1, -1,  1, -1,  1,  1, -1, -1,  1,  1, -1,  1, -1, -1, -1,  1, -1,
       -1, -1, -1, -1, -1, -1, -1, -1,  1,  1, -1, -1,  1, -1, -1],
      dtype=int64)
```

Accuracy and score for new_train.csv

	precision	recall	f1-score	support
-1	0.71	0.84	0.77	153
1	0.80	0.65	0.71	147
accuracy			0.75	300
macro avg	0.76	0.74	0.74	300
weighted avg	0.75	0.75	0.74	300

```
knn.score(Xv_test,y_test)
```

```
0.7466666666666667
```

Predicting Sentiment for Test Dataset (new_test.csv)

```
result[:100]
```

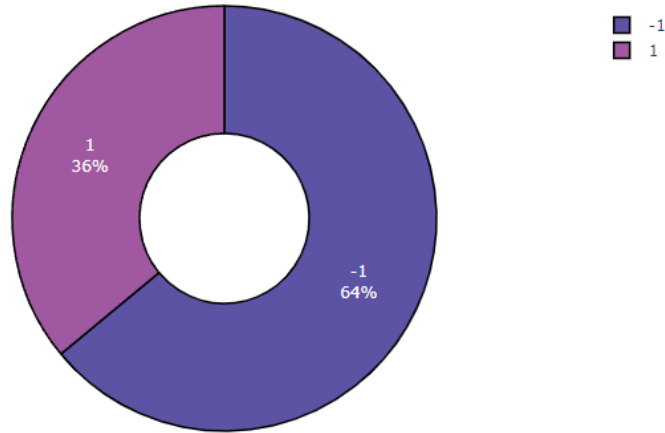
```
array([-1,  1, -1,  1, -1,  1, -1,  1,  1,  1, -1, -1, -1, -1,  1, -1,
        1,  1, -1, -1, -1, -1, -1, -1,  1, -1, -1,  1, -1,  1, -1,  1,
       -1, -1,  1,  1,  1, -1, -1, -1, -1,  1, -1, -1,  1,  1, -1,  1, -1,
       -1, -1, -1,  1, -1, -1, -1, -1, -1, -1, -1,  1, -1, -1, -1, -1,  1,
        1, -1,  1, -1, -1, -1, -1,  1, -1, -1,  1, -1, -1,  1, -1,  1, -1,
        1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1],
      dtype=int64)
```

The final prediction for new_test.csv is stored in Predicted_test.csv

Delimiter:

	Text	sentiment
1	got takeout last night horrible something must happened people working disorganized checked food left ordered fajitas littl...	-1
2	girls sweet prices reasonable stand bed hot make sure adjust time might super white lol	1
3	rudest people ever encountered husband wife owned business called service wife unbelievably unnecessarily rude demandi...	-1
4	airport coveted destination leads fliers view strip thing redeems hotel tacky slot machines center terminal one thing dont ev...	1
5	last months shown steady decline pisspoor service shall startnfor two weeks row count ondemand crapping friday evening...	-1
6	excellente bouffe pas trop cher luegumes et herbes de qualituies portions guenuereuses souvent le restaurant se remplis rap...	1
7	phoenician dated property need facelift spend money elsewhere	-1
8	sunday afterchurch destination phenomenal hummus pitas pizza wraps like hummus trio grilled chicken add etra pitas cucu...	1
9	happy hour pm slushies midnight shakes	1
10	hurry try ginger cinnamon latte added carmel sauce soy milk ice large whopping oz size delish flavor combination amazing ...	1
11	back january gall bladder removed unepected scary surgery someone like ever wisdom teeth referred dr tengs office friend ...	-1
12	total loss bread like wonder bread pasta good sauce lacked flavor price portion minuscule shouldnt go away hungry dollar it...	-1
13	horrible service gave food another table	-1
14	lets get fired claps spirit sprinkles sorry echeerleader couldnt help think chant walked p okay talk joint first thing came mind...	-1
15	overpriced nothing special nmy margarita came small cafeteriastyle glass barely liquor definitely enough get lit beer helped ...	-1
16	anyone says phoeni lacks culture yet visit symphony hall symphony hall reflects sense arizona culture sense architecture re...	1
17	airport need update lack luster healthy options food although common airports still random oygen bars people still use bat...	-1
18	first service pretty good guys nice helpful ordered garlic knots tasty however came inside check knots saw gentalmen finishi...	1
19	three times wife kids twice work buddies times food ecellent service top notch atmosphere finenvisit camarones special reco...	1
20	absolute worst nail salon ive ever never go would recommend place anyone nails never looked awful started come within w...	-1
21	skin place on another date time dont time would recommend nathans hotdog something trashonthe place reaked old beer w...	-1

Frequency of positive and Negative reviews in Test dataset



Conclusion

During creation of this project we learned about how a KNN model can be implemented in the python language. We also learned how preprocessing works and how to extract the data and fit it to the machine learning model.

We also learned how to perform sentiment analysis using machine learning techniques. These techniques can be applied to real life scenarios where businesses and users can use them to identify the popularity, quality and other various aspects of the service/product available on the platform.

References

- [1] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, New Avenues in Opinion Mining and Sentiment Analysis, IEEE Intelligent Systems, 28(2), 2013, pp 15-21. <https://doi.org/10.1109/MIS.2013.30>.
- [2] Sasikumar.A.N, “Sentimental Analysis of Social Networking Sites for Categorization of Product Reviews”, International Journal of Pure and Applied Mathematics, Vol. 117, 2017, pp. 87 – 92.
- [3] J. Mannar Mannan, J, Jayavel, “An adaptive sentimental analysis using ontology for retail market”, IJET, Vol.7, No 1.2, 2018.
- [4] V. Uma. Ramya, K. Thirupathi Rao, “Sentiment Analysis of movie review using Machine Learning techniques", IJET, Vol.7 (2.7), 2018.