Sentiment Analysis

Venkata Reddy

statinfer

Content

- What is sentiment analysis
- Data preparation
- Bayes method
- Naïve Bayes method for finding sentiments
- Accuracy and Confusion matrix

statinfer

What is sentiment Analysis

- Is the service review positive or negative?
- Positive and negative responses in a survey verbatim
- In our given context is that statement positive or negative ?
- Is that blog post positive or negative?
- How are people writing reviews for a movie? Positively or negatively?

- Also Known As
  - Opinion mining
  - Sentiment mining
  - Verbatim Analysis
  - Subjectivity detection

Issues with sentiment analysis

- **This vacuum cleaner sucks**
- **I never had such pizza before, not sure about future either**
- No action, no drama, no comedy, no romance, just pure horror
- The food was not good, it was bad
  The food was not bad, it was good

Limitations of Finding Sentiments

- Text data itself is unstructured / semi structured
- Sarcasm is very difficult to understand
- Sometimes training data doesn't have any strong opinion. Neutral statements
- Strong short documents are often overshadowed by large individual documents

Generic vs Customised algorithm

- Generic text mining algorithms might not work well on all types of text data
- We need to train the model/ dictionary/ DTM to make the model accurate

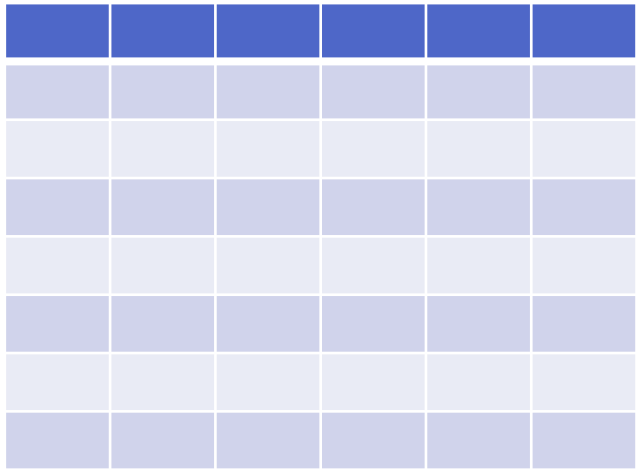
- You will have low accuracy with generic parameters

Two Steps in NLP Model building

Step1=> Convert text data into numerical data

Step2=> Build models on numerical data – Sentiment Analysis model

Step1



Step2

Case Study: Twitter Sentiment Analysis

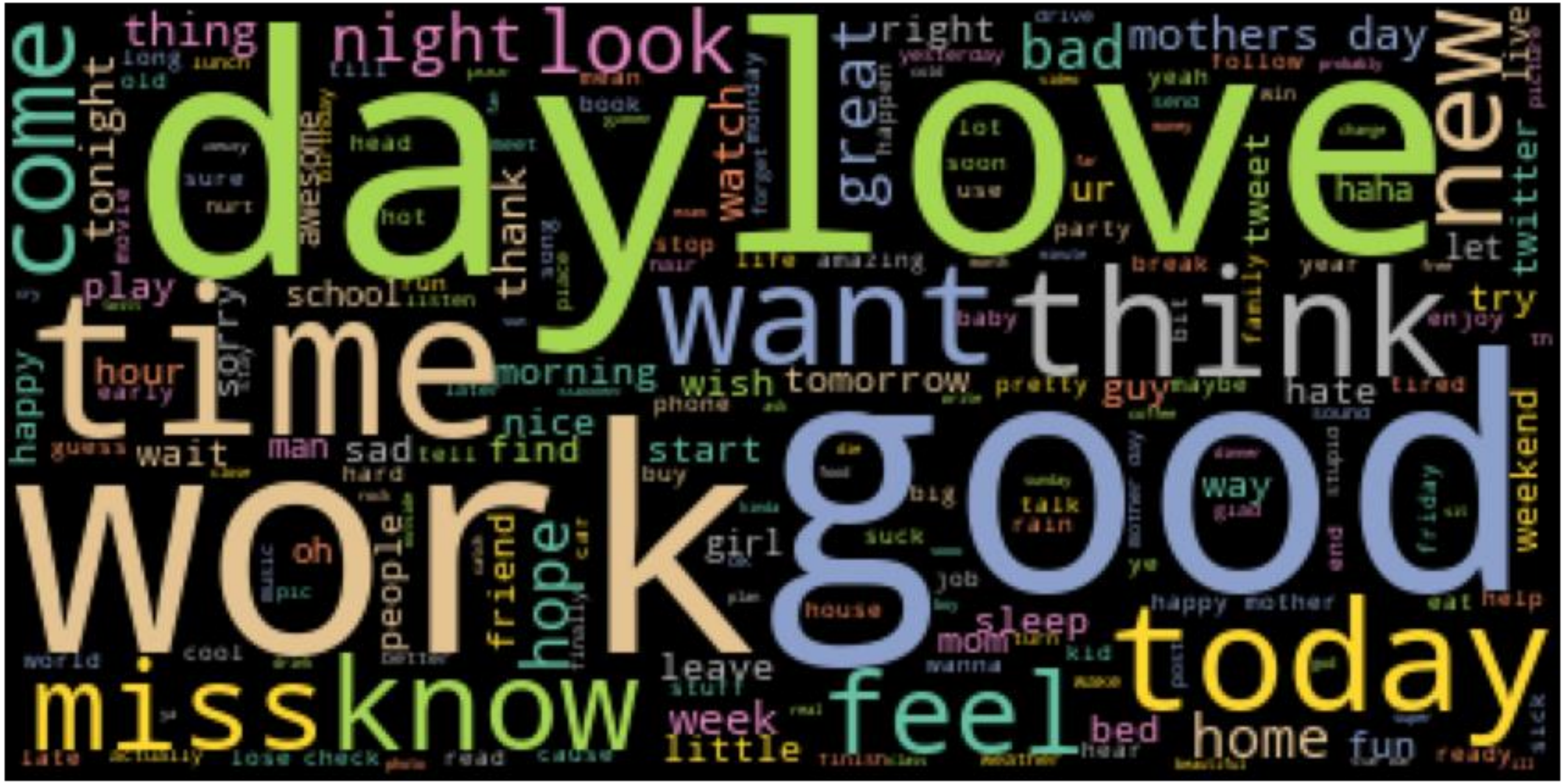
Import data and pre-process

```
twitter_data=pd.read_csv("https://raw.githubusercontent.com/venkatareddykonasani/Datasets/master/Twitter_Sentiment/Twitter_Sentiment_Data.csv")

twitter_data.sample(10)

pre_processing(input_data=twitter_data, text_col="raw_tweet")
```

Word Cloud


day
love
good
work
time
think
want
today
miss
know
feel
new
come
look
night
thing
right
bad
mothers day
great
watch
tonight
thank
home
fun
week
little
leave
bed
sleep
mom
people
hope
friend
oh
hour
happy
wait
sad
man
find
nice
start
morning
wish
tomorrow
guy
hate
play
school
sorry
awesome
tweet
twitter
haha
let
try
enjoy
weekend
way
girl
lot
soon
use
ur
party
break
year
live
head
hot
book
mean
till
old
long
lunch
sure
hurt
stop
life
amazing
baby
pretty
tired
phone
hard
buy
tell
guess
world
cool
late
lose
check
read
cause
help
eat
ready
happy mother
house
job
talk
rain
suck
big
kid
wanna
hear
follow
yeah
send
yesterday
drive

statinfer

Document Term Matrix

	aaaah	aah	abandon	ability	abit	able	absolutely	abt	ac	academy	accept	access	accident	a
0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
27476	0	0	0	0	0	0	0	0	0	0	0	0	0	
27477	0	0	0	0	0	0	0	0	0	0	0	0	0	
27478	0	0	0	0	0	0	0	0	0	0	0	0	0	
27479	0	0	0	0	0	0	0	0	0	0	0	0	0	
27480	0	0	0	0	0	0	0	0	0	0	0	0	0	

27481 rows × 3474 columns

What is Naïve Bayes model

Bayes Theorem

Bayes theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event

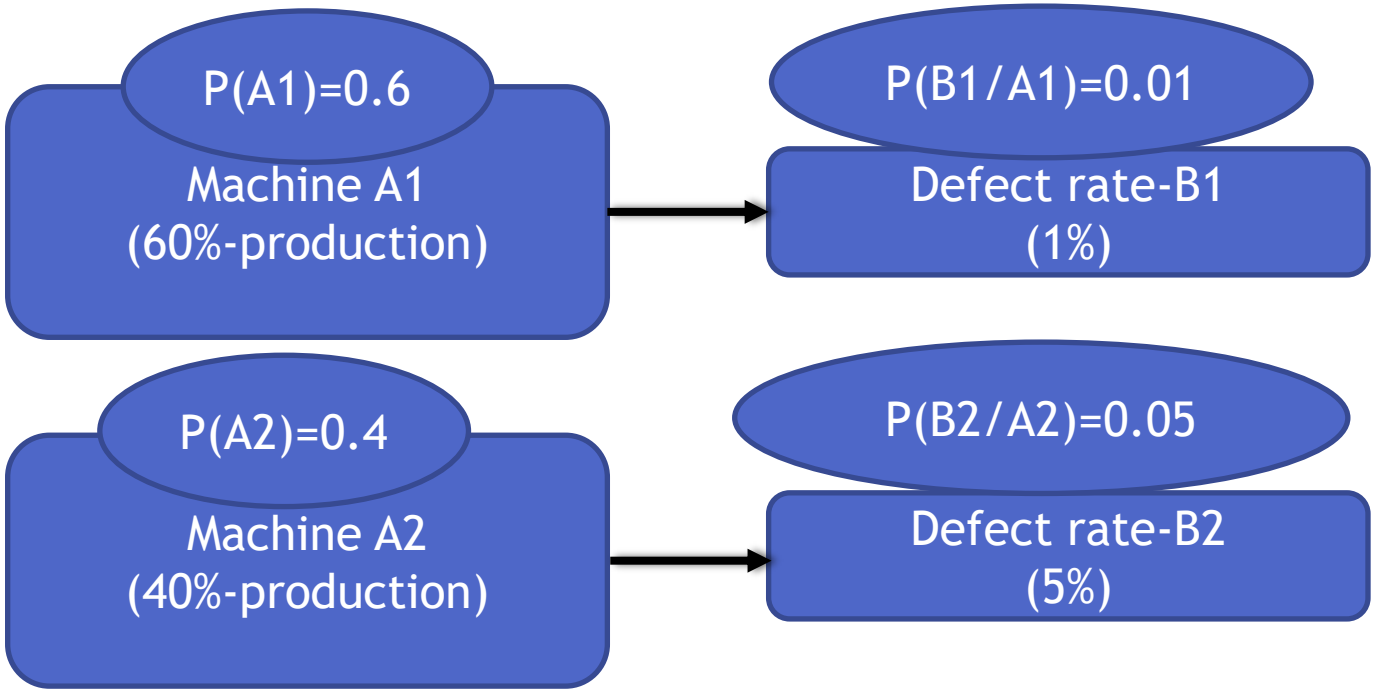
$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)},$$

where A and B are events and $P(B) \neq 0$.

- $P(A)$ and $P(B)$ are the probabilities of observing A and B without regard to each other.
- $P(A \mid B)$, a conditional probability, is the probability of observing event A given that B is true.
- $P(B \mid A)$ is the probability of observing event B given that A is true.

Understanding Bayes Theorem

In a factory two machines produce bolts. Given a faulty bolt, what is the probability that it is produced by machine-1

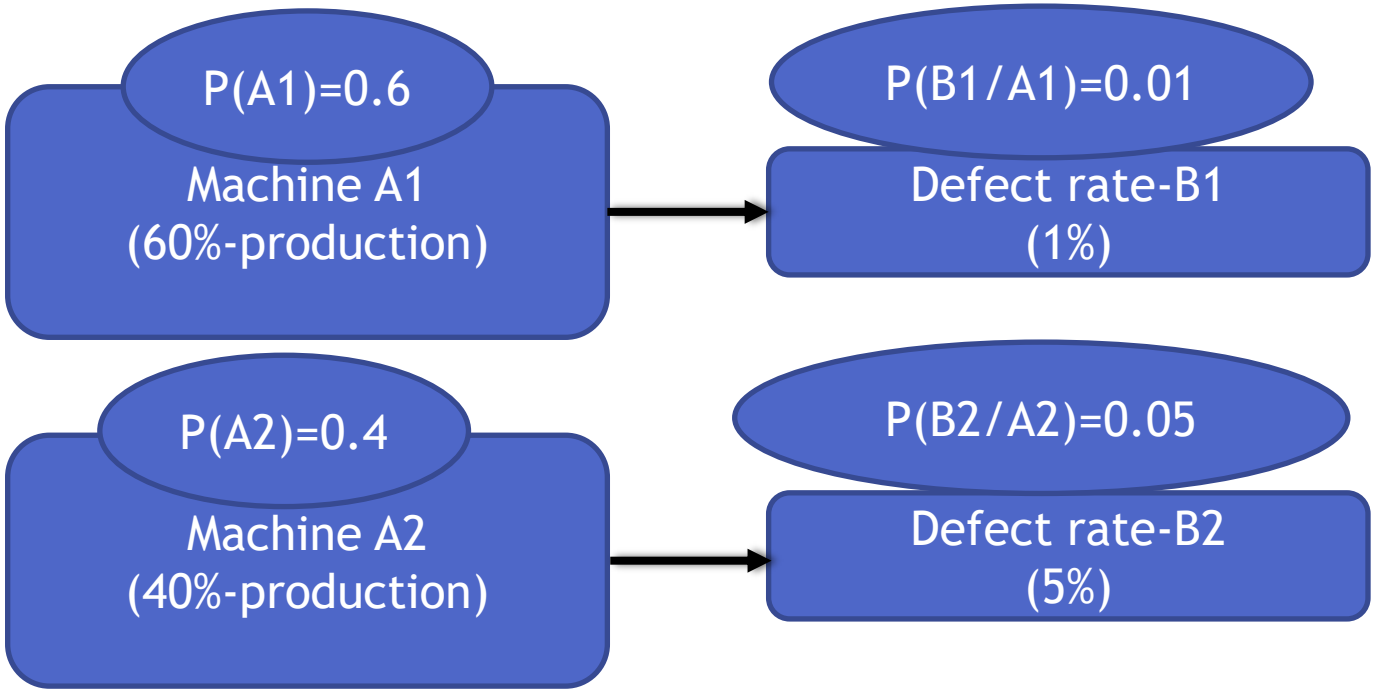


Overall Defect Rate P(B)=
P(B1/A1)*P(A1) + P(B2/A2)*P(A2)

(0.01)*0.6 + (0.05)*(0.4)
0.026

Understanding Bayes Theorem

Given item is already defective; what is the of A1 producing it?



0.026

Understanding Bayes Theorem

- Overall defect percentage is 0.026, take that as final reference. What proportion of 0.026 is taken by M1 and what portion M2 takes

- Overall defective =Weighted Defectives from M1 + Weighted Defectives from M2

- 0.026 = (0.01)*0.6 + (0.05)*(0.4)
- P(B) =P(B1/A1)*P(A1) + P(B2/A2)*P(A2)

Understanding Bayes Theorem

- Overall defect percentage is 0.026, take that as final reference. What proportion of 0.026 is taken by M1 and what portion M2 takes

- Overall defective =Weighted Defectives from M1 + Weighted Defectives from M2

- 0.026 = (0.01)*0.6 + (0.05)*(0.4)
- P(B) =P(B1/A1)*P(A1) + P(B2/A2)*P(A2)

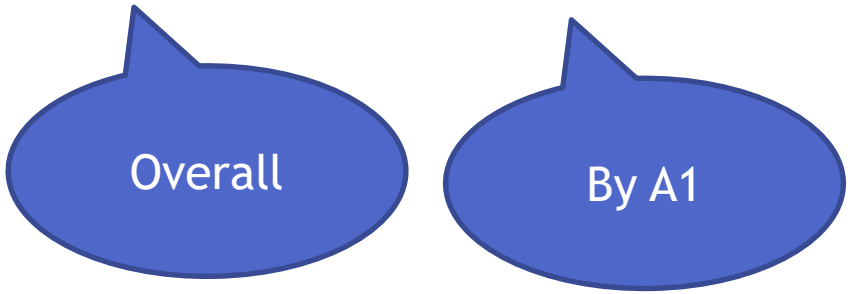
- Given item is already defective; what is the chance of A1 producing it?
- What is A1's contribution in overall?

Understanding Bayes Theorem

- Overall defect percentage is 0.026, take that as final reference. What proportion of 0.026 is taken by M1 and what portion M2 takes

- Overall defective =Weighted Defectives from M1 + Weighted Defectives from M2

- 0.026 = (0.01)*0.6 + (0.05)*(0.4)
- P(B) =P(B1/A1)*P(A1) + P(B2/A2)*P(A2)



- Given item is already defective; what is the chance of A1 producing it?
- What is A1's contribution in overall?

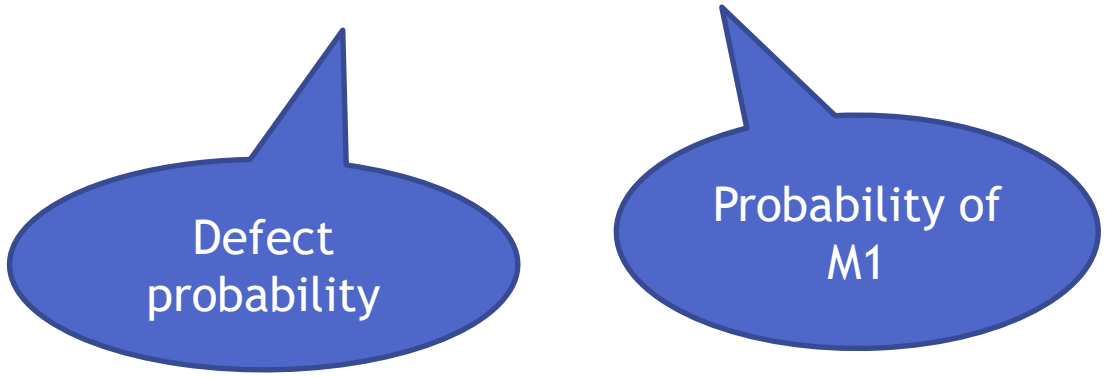
Understanding Bayes Theorem

- Overall defect percentage is 0.026, take that as final reference. What proportion of 0.026 is taken by M1 and what portion M2 takes

- Overall defective =Weighted Defectives from M1 + Weighted Defectives from M2

Given item is already defective

- P(B)/ P(B) =(P(B1/A1)*P(A1))/ P(B) + (P(B2/A2)*P(A2))/ P(B)



statinfer

Understanding Bayes Theorem

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)},$$

Event A is producing using machine
Event B is defective item production

where A and B are events and $P(B) \neq 0$.

- $P(A)$ and $P(B)$ are the probabilities of observing A and B without regard to each other.
- $P(A \mid B)$, a conditional probability, is the probability of observing event A given that B is true.
- $P(B \mid A)$ is the probability of observing event B given that A is true.

1= (P(B1/A1)*P(A1))/ P(B) + (P(B2/A2)*P(A2))/ P(B)

Understanding Bayes Theorem

- Given a bolt is defective what is the probability that it is coming from a particular machine

- Given a new document what is the probability that it is coming from positive set / negative set

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Naïve Bayes theorem for sentiment analysis

statinfer

- New document *d*;
- Classes={c_1,c_2}
- Compute the Bayes probability that *d* is in each class *c* $\in$C

$$\Pr(c_i|d) = \frac{\Pr(d|c_i)\Pr(c_i)}{\Pr(d)}$$

- P(d) -> probability of words in a specific document, across all docs
- P(d/c) -> Probability of words in a specific class
- P(c) -> Probability of a class

Naïve Bayes theorem for sentiment analysis

statinfer

- New document *"Awesome"*;
- Classes={Positive, Negative}
- Compute the Bayes probability that *"Awesome"* is in each class $c \in C$

$$\Pr(\text{Positive}|\textit{Awesome}) = \frac{\Pr(\textit{Awesome}|\text{Positive})\Pr(\text{Positive})}{\Pr(\textit{Awesome})}$$

- P(d) -> probability of words in a specific document, across all docs
- P(d/c) -> Probability of words in a specific class
- P(c) -> Probability of a class

Naïve Bayes theorem for sentiment analysis

statinfer

- New document *"Awesome"*;
- Classes={Positive, Negative}
- Compute the Bayes probability that *"Awesome"* is in each class $c \in C$

$$\Pr(\text{Positive}|\textit{Awesome}) = \frac{\Pr(\textit{Awesome}|\text{Positive})\Pr(\text{Positive})}{\Pr(\textit{Awesome})}$$

$$\Pr(\text{Positive}|\textit{Awesome}) = \frac{\left(\frac{800}{1000}\right) * \left(\frac{1000}{2}\right)}{\left(\frac{1000}{2000}\right)}$$

- P(d) -> probability of words in a specific document, across all docs
- P(d/c) -> Probability of words in a specific class
- P(c) -> Probability of a class

Finally

- Naïve Bayes method gives us the positive or negative sentiment of a given document

Train and Test Data

```
dtm_v1['sentiment_label']=twitter_data['sentiment_label']

#remove neutrals
dtm_v1=dtm_v1[dtm_v1['sentiment_label'] != "neutral"]
print(dtm_v1['sentiment_label'].value_counts())

X=dtm_v1.drop(['sentiment_label'], axis=1)
y=dtm_v1['sentiment_label']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

statinfer

Model Building

```
from sklearn.naive_bayes import MultinomialNB
senti_model = MultinomialNB()
#Fitting model to our data
senti_model.fit(X_train, y_train)

print("Train Accuracy", senti_model.score(X_train,y_train))
print("Test Accuracy", senti_model.score(X_test,y_test))
```

```
print("Train Accuracy", senti_model.score(X_train,y_train))
print("Test Accuracy", senti_model.score(X_test,y_test))
```

```
Train Accuracy 0.8928953399541635
Test Accuracy 0.84967919340055
```

statinfer

New Review Prediction

```
t1 ="Awesome experience. Go for it. It is a great place"
t2 ="Very bad day for me today. I would like to forget it as soon as possible"
t3 = "I am the way i am. If I wasn't what ever you say i am. because I am the way I am"

tweet_list=[t1,t2,t3]

new_comment= pd.DataFrame({"text":tweet_list})

#Spelling Correction
from textblob import TextBlob
new_comment["text_corrected"]=new_comment["text"].apply(lambda x:"".join(TextBlob(x).co
rrect()))
pre_processing(input_data=new_comment, text_col="text_corrected")
```

New Review Prediction

```
countvec = CountVectorizer()
dtm_newcomment = pd.DataFrame(countvec.fit_transform(new_comment['text_col_clea
n']).toarray(), columns=countvec.get_feature_names(), index=None)
#print(dtm_newcomment)

dtm_v2=dtm_v1.drop(["class"],axis=1)
dtm_newcomment_final=pd.DataFrame(columns=dtm_v2.columns.values)
dtm_newcomment_final=dtm_newcomment_final.append(dtm_newcomment)
dtm_newcomment_final=dtm_newcomment_final.fillna(0)

print("****Make sure that New DTM and old DTM have same number of columns***")
print("New DTM Shape", dtm_newcomment_final.shape)
print("Overall DTM Shape",dtm_v2.shape)
```

New Review Prediction

text	Sentiment
Awesome experience. Go for it. It is a great p...	positive
Very bad day for me today. I would like to for...	negative
I am the way i am. If I wasn't what ever you s...	positive

Example-2 – Amazon Yelp reviews

Amazon_yelp reviews data

- Download the dataset
- Pre-process it
- Build a sentiment analysis model
- Try to get predictions for new reviews

Conclusion

Conclusion

- Data cleaning is critical
- Naïve Bayes is most widely used method in sentiment analysis
- It can be used in document categorization as well

Thank you