# Sentiments in Twitter

*Kaneesha Dawood*

# The problem

## Company

Twitter is one of the popular micro-blogging platforms that provide data for a wide range of users.

## Context

Tweets come from a variety of sources (consumers, companies, public officials etc). Consumers use Twitter to connect to their brand. Likewise, their sentiments are an invaluable asset to the company.

## Problem statement

**"How can a company use Twitter sentiments to predict consumer spending patterns?"**

# Problem Identification & Approach

Goal: Identify the positive and negative sentiments in making a purchasing decision.

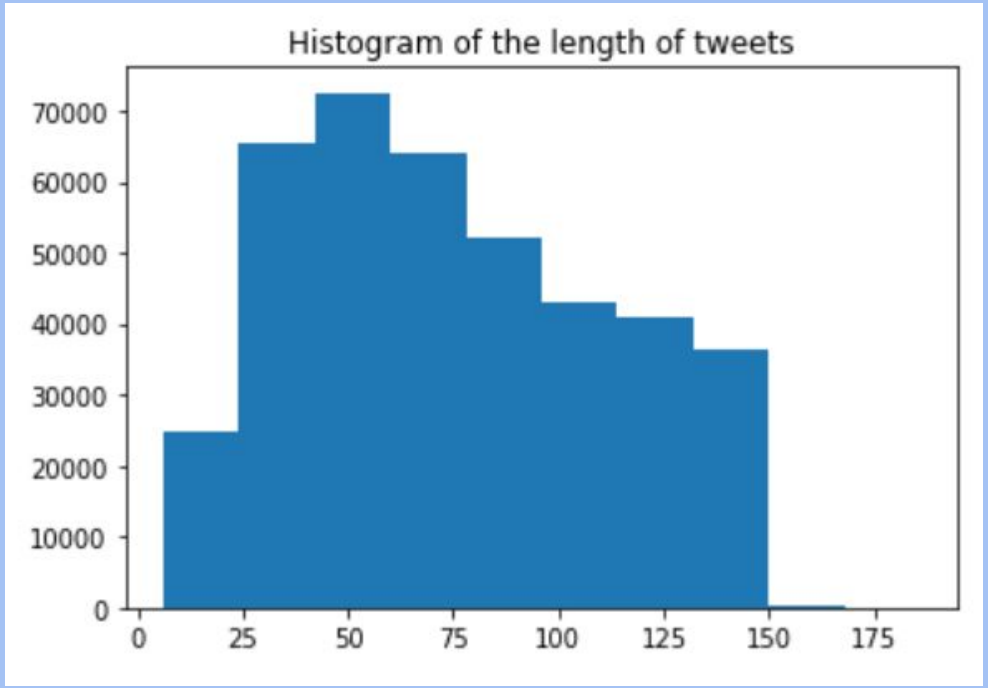
*Approach*

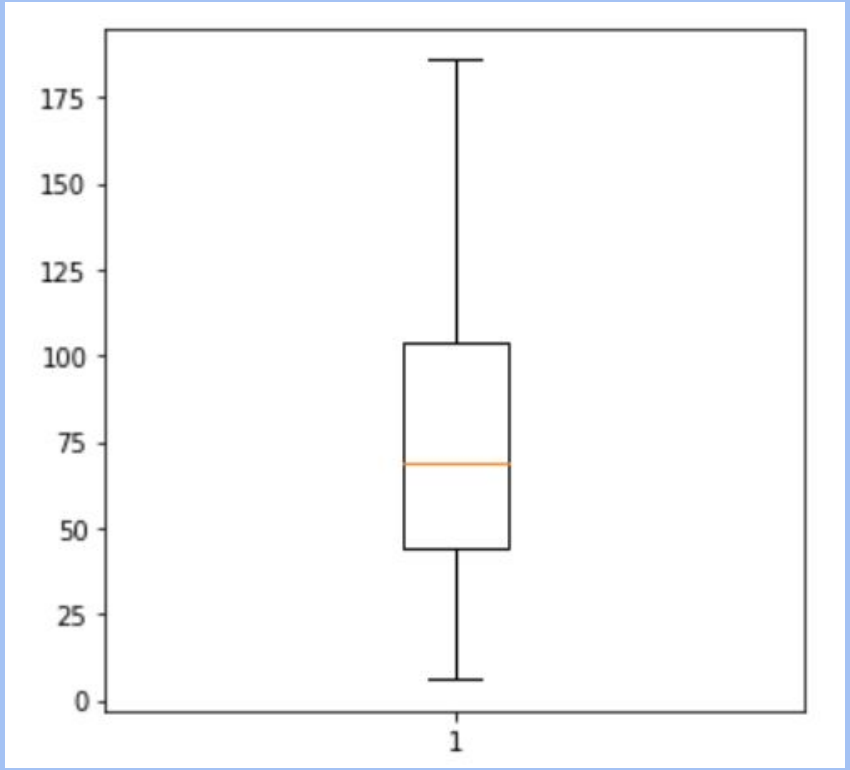
1. Read the Twitter Data
2. Data Exploration
3. Data Cleaning and Preprocessing
4. Exploratory Analysis and Visualization
5. Feature Extraction
6. Model Building and Evaluation

# Twitter Dataset

| | sentiment | user_id | date | query | user | tweet |
|---|---|---|---|---|---|---|
| 0 | 0 | 1467810369 | Mon Apr 06 22:19:45 PDT 2009 | NO_QUERY | _TheSpecialOne_ | @switchfoot http://twitpic.com/2y1zl - Awww, t... |
| 1 | 0 | 1467810672 | Mon Apr 06 22:19:49 PDT 2009 | NO_QUERY | scotthamilton | is upset that he can't update his Facebook by ... |
| 2 | 0 | 1467810917 | Mon Apr 06 22:19:53 PDT 2009 | NO_QUERY | mattycus | @Kenichan I dived many times for the ball. Man... |
| 3 | 0 | 1467811184 | Mon Apr 06 22:19:57 PDT 2009 | NO_QUERY | ElleCTF | my whole body feels itchy and like its on fire |
| 4 | 0 | 1467811193 | Mon Apr 06 22:19:57 PDT 2009 | NO_QUERY | Karoli | @nationwideclass no, it's not behaving at all.... |
| 5 | 0 | 1467811372 | Mon Apr 06 22:20:00 PDT 2009 | NO_QUERY | joy_wolf | @Kwesidei not the whole crew |
| 6 | 0 | 1467811592 | Mon Apr 06 22:20:03 PDT 2009 | NO_QUERY | mybirch | Need a hug |
| 7 | 0 | 1467811594 | Mon Apr 06 22:20:03 PDT 2009 | NO_QUERY | coZZ | @LOLTrish hey long time no see! Yes.. Rains a... |
| 8 | 0 | 1467811795 | Mon Apr 06 22:20:05 PDT 2009 | NO_QUERY | 2Hood4Hollywood | @Tatiana_K nope they didn't have it |
| 9 | 0 | 1467812025 | Mon Apr 06 22:20:09 PDT 2009 | NO_QUERY | mimismo | @twittera que me muera ? |

# Univariate Analysis



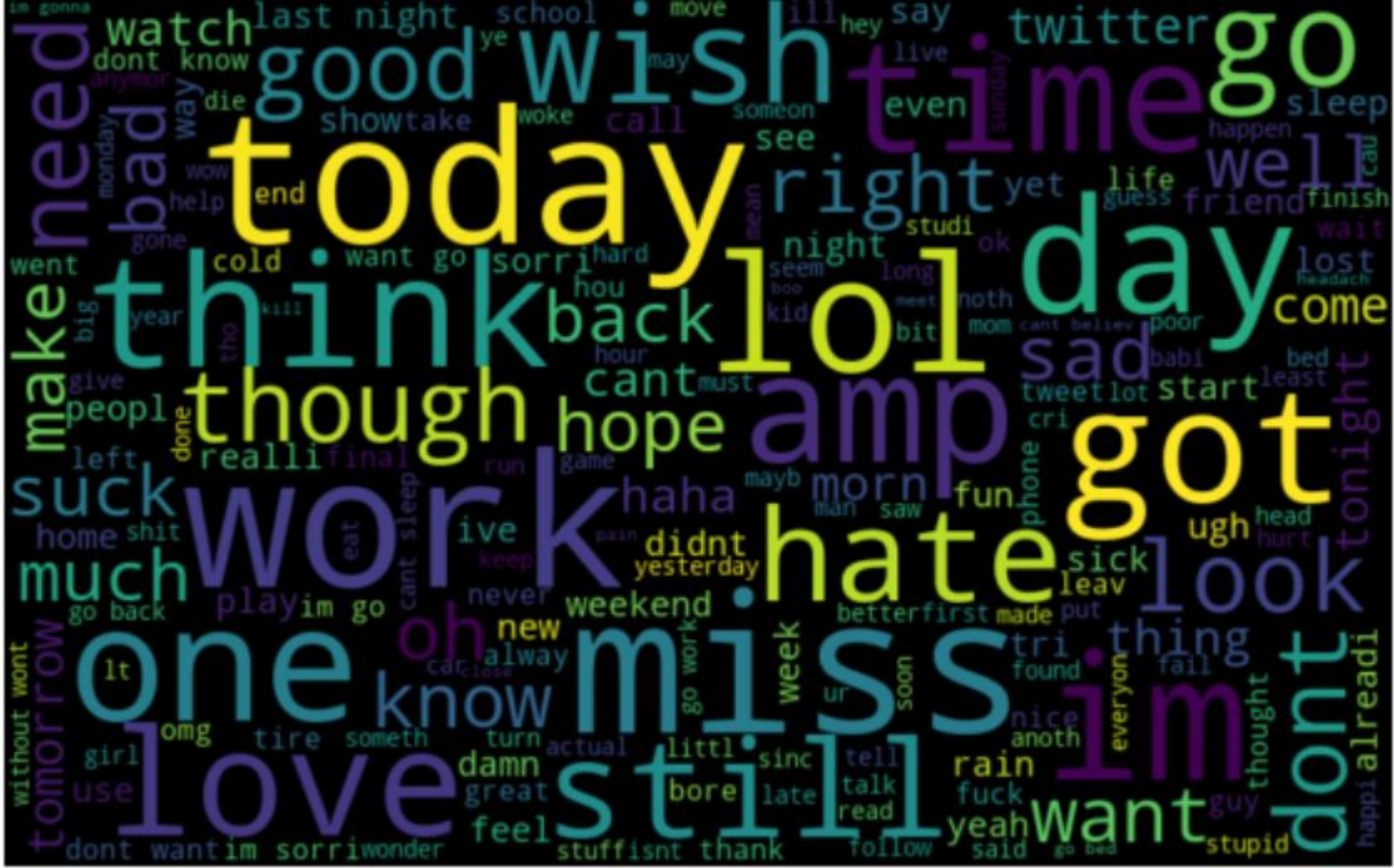



# Data Preprocessing and Cleaning

**The raw Twitter text include people's casual opinions in the form of informal language which add noise to our data.**

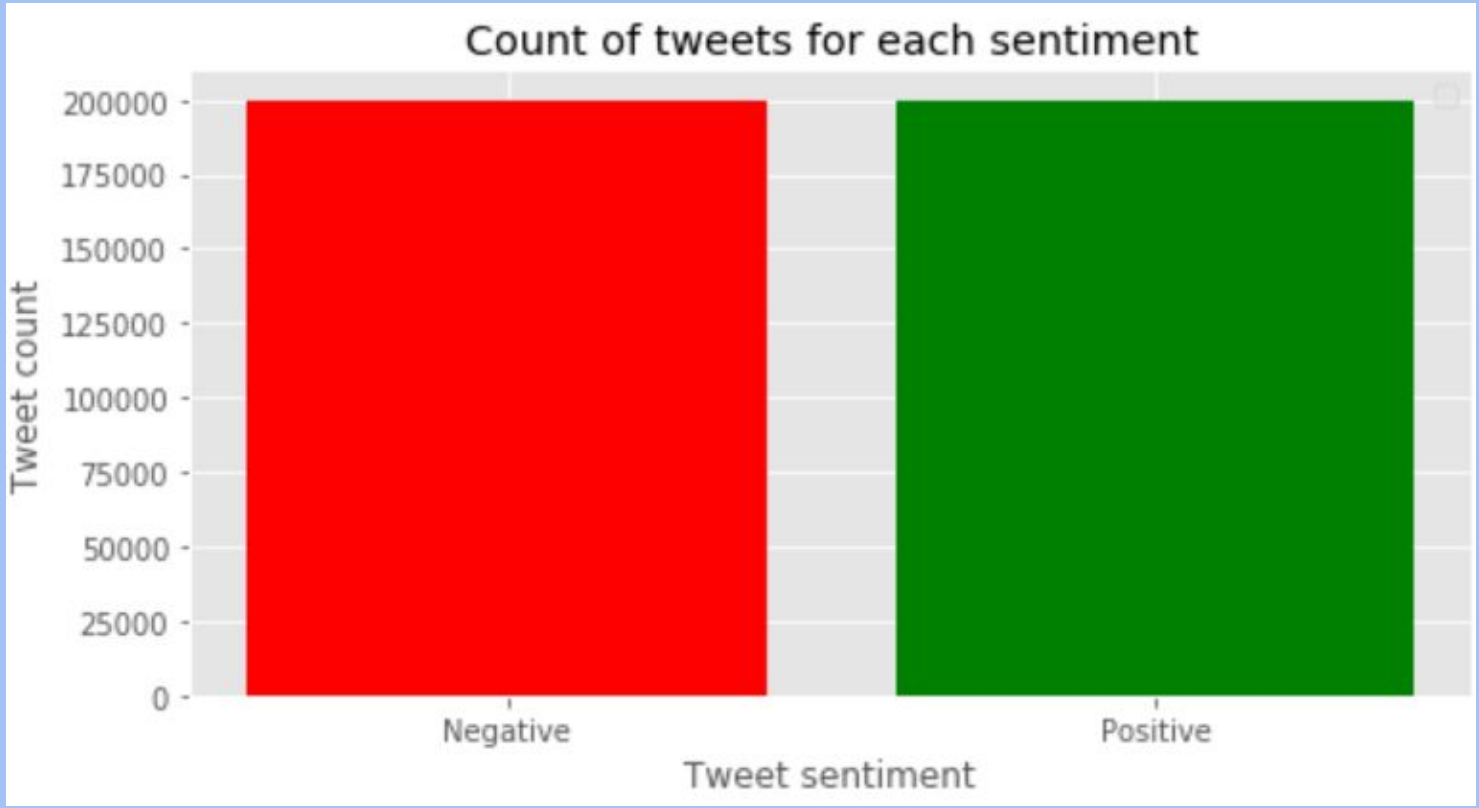
**Steps to Preprocess:**

1. **Drop Columns ‘user_id’, ‘date’, ‘query’, ‘user’**
2. **Convert tweet to lower case**
3. **Remove punctuation, numbers, and irrelevant characters**
4. **Eliminate extra white space**
5. **Tokenization**
6. **Remove Stop words**
7. **Lemmatization**
8. **Stemming**



im gonna
last night
school
move
ill
hey
say
twitter
go
watch
dont know
ye
may
live
need
good
wish
time
anymor
die
way
bad
someon
even
sleep
showtake
woke
call
sunday
happen
monday
wow
see
well
today
right
yet
life
help
end
guess
friend
finish
studi
mean
wait
gone
ok
went
cold
want go
sorri
hard
night
lost
headach
think
lol
day
hou
seem
long
big
year
kill
boo
meet
noth
come
make
back
kid
mom
cant believ
poor
tho
bit
sad
hour
babi
bed
give
cant
must
least
though
amp
tweet
lot
start
peopl
hope
tonight
done
cri
left
realli
final
got
run
game
phone
mayb
work
suck
haha
morn
fun
man
saw
head
home
shit
eat
pain
ugh
hurt
cant sleep
ive
didnt
much
hate
sick
keep
yesterday
look
leav
never
play
im go
weekend
betterfirst
made
put
go back
oh
new
tomorrow
week
thing
tri
without
wont
alway
lt
car
close
go work
miss
fail
one
know
im
found
soon
ur
everyon
alreadi
dont
nice
omg
tire
someth
turn
actual
littl
anoth
thought
love
girl
sinc
tell
damn
still
rain
use
great
bore
late
talk
happi
guy
fuck
read
feel
want
yeah
dont want
im sorri
wonder
stuffisnt
thank
follow
said
go bed
stupid

# Twitter Sentiments



# Train-Test Split

Identify the Variables

- Y = target variable = sentiment
- X = features = tweets

Data is distributed into two sets on a 75:25 ratio

- Training set
- Testing set

```
Splitting train and test dataset into 75:25
Train data distribution:
1    150219
0    149781
Name: sentiment, dtype: int64
Test data distribution:
0    50219
1    49781
Name: sentiment, dtype: int64
Split complete
```

# Feature Extraction

I used the following methods to extract words from our clean dataset:

- TF-IDF Features (Term Frequency- Inverse Document Frequency):

  To identify and extract the most important words

- Bag-of-Words Features:

  Words are extracted and converted into the binary form

# Model Evaluation

| | Random Forest | Logistic Regression | Naive Bayes | Gradient Boosting |
|---|---|---|---|---|
| **Accuracy** | 98.983% | 98.744% | 90.273% | 86.737% |
| **F1 Score** | 0.99 | 0.99 | 0.91 | 0.88 |

# Best Model

The best classifier in predicting the Twitter Sentiment is the Random Forest Classifier

- Predictive performance: Higher accuracy scores on the Training and Testing data (98%)
- Captures both positive and negative words and predicts the outcome of the future Twitter Sentiment
- Reliable feature importance estimate

```
Training Random Forest Classifier
Predicting the train data
Training accuracy: 98.966%
Predicting the test data
Testing accuracy: 98.983%
Confusion Matrix:
[[49202  1017]
 [    0 49781]]
Classification Report
              precision    recall  f1-score

           0       1.00      0.98      0.99
           1       0.98      1.00      0.99
```

# Conclusion

My goal was to detect the positive and negative sentiments from the Twitter data to help the company predict their consumer spending patterns.

- Preprocessed data using NLP techniques
- Extracted features using two methods
- Built several models to predict performance
- Evaluated models using 2-3 metrics
- Best model: Random Forest Classifier

# Future Work

- Increase the sample size and analyze the impact of model performance
- Use advanced Natural Language Processing techniques to classify the text
- Build more models and evaluate against different performance metrics
- Tune the hyperparameters through cross validation