Emotion Analysis Based Activity Recommendation from Social Media Posts

BY

Kathyayani Bolgam, Amrutha Polu, Modassir Ansari, Abhisekhar

Guide

Dr. Selvi C

Indian Institute of Information Technology Kottayam

Introduction

- Rapid growth in the usage of social media.
- Information is an important basis for inferring user's state.
- The analysed information is crucial in marketing or development point of view.
- In this project, tweets of users in the healthcare domain will be analysed .

Significance of Health Care System

- Core area in day-today life
- Derives insights and predictions of any health condition
- Important platform for healthcare services
- Main objective is to provide
    a. Quality information
    b. Trustworthiness
    c. Authenticity
    d. Privacy
- Derives outcomes such as recommending
    a. Diagnosis
    b. Health insurance
    c. Clinical pathway based treatment methods
    d. Alternative medicines

Indian Institute of Information Technology Kottayam

LITERATURE REVIEW

S.No	Title	Authors	Aim/ Objective	Approach	Advantages	Limitations
1.	Sentiment Analysis of Health Care Tweets: Review of the Methods Used [3]	Sunil Gohil, Sabine Vuik, Ara Darzi	To understand which tools would be available for sentiment analysis of Twitter health care research, by reviewing existing studies in this area and the methods they used. to determine which method would work best in the health care settings, by analyzing how the methods were used to answer specific health care questions, their production, and how their accuracy was analyzed.	The study compared the types of tools used in each case and examined methods for tool production, tool training, and analysis of accuracy	collaborative approach could be used to produce a more advanced and accurate tool for the health care setting using subject-specific lexicons and complementary health care–based features	Any exact method or approach was not proposed for validating accuracy

2.	Semantics-enhanced Recommendation System for Social Healthcare [6]	Nazia Zaman, Juan Li	proposing an effective personalized recommender system which can recommend products, services, or information to users in the healthcare social network to speed their recovery and improve healthcare outcomes. between people in the social network to make recommendations	We enhance the user based collaborative filtering with social semantics that will involve the aforementioned intuitions and address the issue of lack of rating.	focused on features of health-concerns in the user profile.	Facebook, WebMD, LIVESTRONG and Wikipedia, then collected 500 posts, which are categorized under 50 diseases and 50 symptoms. Each user maintains a "post preference" data set

3.	Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes – Support Vector Machine (NBSVM) Classifier [1]	Abbi Nizar Muhammad, Saiful Bukhori, Priza Pandunata	This research would like to give a contribution to the development of sentiment analysis approach using Naïve Bayes - Support Vector Machine (NBSVM) method with a Binary Classification approach. T	Naïve Bayes –Support Vector Machine (NBSVM) Classifier	the performance test results of positive and negative sentiments using Naïve Bayes Support Vector Machine method produces the highest precision, recall and F1 scores Support Vector Machine is extensively used as a basic line in tasks related to texts but the performance varies significantly in all variants, features, and numbers of data collection.	The type of data used is Indonesian text obtained from several sources

4	Patients' and health professionals' use of social media in health care: Motives, barriers and expectations [4]	Marjolijn L. Antheunis a, Kiek Tates a , Theodoor E. Nieboer b	To investigate patient's and health professional's motives and use of social media for health-related reasons and barriers and expectations for health-related social media use.	They conducted a descriptive online survey among 139 patients and 153 health care professionals in obstetrics and gynecology. In this survey, they asked the respondents about their motives and use of social network sites.	Patients use 31.7% health related social media compared to 99.3% personal use of social media, whereas health professionals use 26.8% of professional use in comparison to 59.3% of personal use of social media.	did not register the number of years of experience of the health professionals. It can be argued, that younger professionals are keener to use social media for health-related reasons than their older colleagues Since most of the participants were found online, there may have been a bias favoring patients who are more "e-ready."

5	VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text [5]	C.J. Hutto , Eric Gilber	They introduce VADER, a straightforward rule-based model for general sentiment analysis, and evaluate its performance against eleven typical state-of-the-art benchmarks, such as SentiWordNet, LIWC, ANEW, the General Inquirer, and machine learning-focused methods using Naive Bayes, Maximum Entropy, and Support Vector Machine algorithms.	They first create and empirically validate a gold-standard collection of lexical features that are adjusted to sentiment in microblog-like contexts using a combination of qualitative and quantitative methodologies. Then, taking into account five overarching principles that represent grammatical and syntactic patterns for expressing	it is both quick and computationally economical without sacrificing accuracy. Running directly from a standard modern laptop computer with typical, moderate specifications, a corpus that takes a fraction of a second to analyze with VADER can take hours when using more complex models like SVM. The lexicon and rules used by VADER are directly accessible, not hidden within a machine-access-only black-box. VADER is therefore easily inspected	For simple binary (positive/negative) classification on single sentences, the state-of-the-art accuracy is approximately 80%, while for the more challenging multi-class situation that includes a third (neutral) class, accuracy tends to hover in the 60% range for social media writing.

				reinforcing feeling intensity, They integrate these lexical elements. Intriguingly, they find that VADER outperforms individual human raters and generalizes more favorably across contexts than any of other benchmarks when used to evaluate the sentiment of tweets using our sparse rule-based model.	extended or modified. Anyone familiar with LIWC(LIWC is a transparent text analysis program that counts words in psychologically meaningful categories.) can also able to use VADAR. not only in the computer science field, but also Sociologists, psy-chologists, marketing researchers, or linguists who are comfortable using LIWC. it does not require an extensive set of training data, yet it performs well in diverse	

Problem Statement

Given a set of tweets T, where

$T_i = \{t_1, t_2, .., t_n\}$, $1 \leq i \leq n$ and a set of activities or suggestions

$A_j = \{a_1, a_2, ., a_m\}$ $1 \leq j \leq m$ for each health condition H.

Our objective is to divide the tweets in to a set of health conditions through text analysis like

$H_k = \{h_1, h_2, ..., h_k\}$ $1 \leq k \leq l$ based on the semantics of the text in tweets such that one or more activities can be suggested for every health condition.

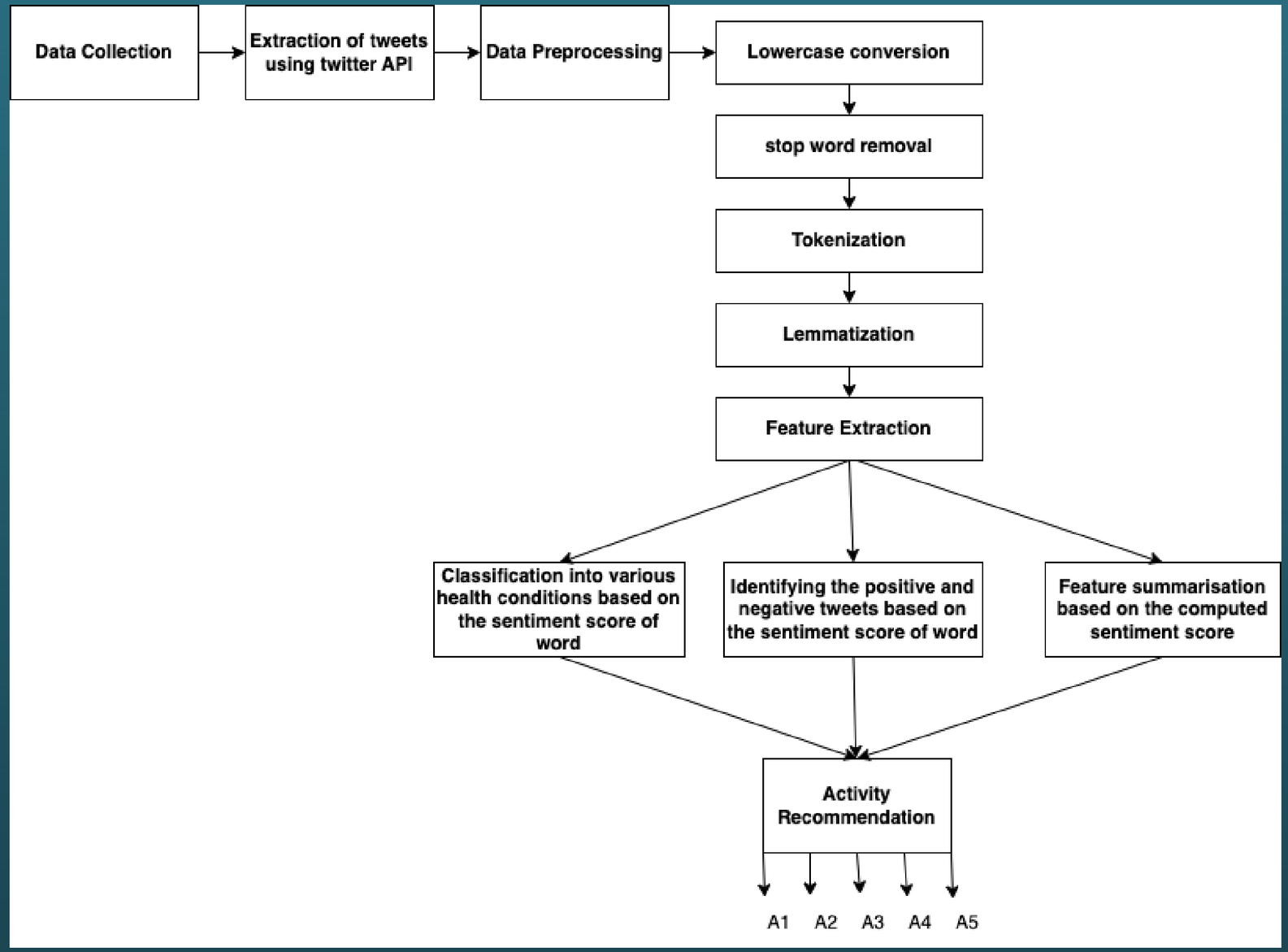
Indian Institute of Information Technology Kottayam

Objective

1. Classifying the tweets based on the calculated sentiment score of the word.
2. Identifying positive and negative tweets based on the calculated sentiment score.
3. Activity recommendation based on feature summarization based on the calculated sentiment score.

Indian Institute of Information Technology Kottayam

ARCHITECTURE


Data Collection
Extraction of tweets using twitter API
Data Preprocessing
Lowercase conversion
stop word removal
Tokenization
Lemmatization
Feature Extraction
Classification into various health conditions based on the sentiment score of word
Identifying the positive and negative tweets based on the sentiment score of word
Feature summarisation based on the computed sentiment score
Activity Recommendation
A1
A2
A3
A4
A5

Indian Institute of Information Technology Kottayam

Experiment Setup

Data Collection using twitter API

```python
import tweepy
import csv
access_token = ""
access_token_secret = ""
consumer_key = ""
consumer_secret = ""

auth = tweepy.auth.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

api = tweepy.API(auth)
csvFile = open('tweets_btp.csv', 'a')

#Use csv writer
csvWriter = csv.writer(csvFile)
csvWriter.writerow(['id', 'Time','Tweet', 'Location'])
for tweet in tweepy.Cursor(api.search_tweets, q = "flu OR cancer OR thyroid OR migraine OR diabetes", count = "500").items():

    # Write a row to the CSV file. I use encode UTF-8
    csvWriter.writerow([tweet.user.id, tweet.created_at, tweet.text.encode('utf-8'), tweet.user_location])
    print(tweet.user.id, tweet.created_at, tweet.text, tweet.user_location)
csvFile.close()
```

Data set obtained

tweets_btp1

id	Time	Tweet	Location
312322346	2022-11-08 10:14:36+00:00	b"Let's be honest; day two of a phase 1 cancer trial never looked this good. #cancer #thyroidcancer #thyroid https:/	London
42296887	2022-11-08 10:15:07+00:00	b'Pembrolizumab: English Drugs Body Recommends Life-Saving Cancer Medicine https://t.co/1m9sgJpV7e https://	Florida
1356818881063473157	2022-11-08 10:15:05+00:00	b"RT @nettybgoode: Well, @ServicesGovAU are at it again. 2 days after my mum finished radiation treatment, I've had to drag her into Centreli\xe2\x80\	
1570261324554293249	2022-11-08 10:15:04+00:00	b'@AlHosnApp where we can take session flu vaccine please guide'	
4836684705	2022-11-08 10:15:03+00:00	b"RT @DFisman: Again, your immune system didn't go all flabby because there's was no flu season in 2021. Your ir	Scotland, United Kingdom
1538053960367497216	2022-11-08 10:15:03+00:00	b'RT @Mary_Collier_AZ: Cancer, wheel of fortune! Predictions Money,Career,Love,Secrets. November Oracle Readings\n-> https://t.co/PSn6hbLtbk	
1154134940323328010	2022-11-08 10:15:03+00:00	b'RT @myESMO: #ESMOTAT23: @plummer_ruth on why you should submit your abstract & help shape the futu	London
1045487782976770048	2022-11-08 10:15:02+00:00	b'RT @CrapulaLeonardo: J\xc3\xa1 falei in\xc3\xbameras vezes, mas volto a repetir: o grande mal, o grande c\xc3\x	São José do Rio Preto, Brasil
520701453	2022-11-08 10:15:01+00:00	b'RT @Ro_Arauz: Mi madre no ha podido luchar. No la han dejado pelear contra el c\xc3\xa1ncer NO diagnosticadc	Barakaldo, España
972574723140878339	2022-11-08 10:15:00+00:00	b'November is Lung Cancer Awareness Month. Recognise the signs and symptoms, find out more at\xe2\x80\xa6 h	West Midlands, England
1160308873	2022-11-08 10:14:58+00:00	b'RT @SKyriakidesEU: Very pleased to address @EconomistEvents #WorldCancerSeriesEurope.\n\nThe public healt	Brussels, Belgium
2498848159	2022-11-08 10:14:58+00:00	b'RT @Sue13800241: Another vile individual who for no reason has decided to reply to my post about bird flu with a	Kent
298484855	2022-11-08 10:14:56+00:00	b'RT @patrianriv: Volvi\xc3\xb3 el C\xc3\xa1ncer, toca seguir sonriendo y seguir adelante. Creo que ahora si vender	VENEZUELA
2960770143	2022-11-08 10:14:56+00:00	b'RT @Annette1Hardy: Well done, @JustStop_Oil. Our consultant surgeon friend is stuck on the M25 right now and	London, England
850644802337112064	2022-11-08 10:14:55+00:00	b'Vitamina C y c\xc3\xa1ncer https://t.co/HB4xrp5asn'	Madrid, Comunidad de Madrid
1519295591850856449	2022-11-08 10:14:55+00:00	b'RT @ClaudioDeLucca: \xc3\x93TIMA NOT\xc3\x8dCIA Vacina experimental contra o c\xc3\xa2ncer de mama gerc	São Paulo, Brasil
33717365	2022-11-08 10:14:55+00:00	b'RT @FOX29philly: Joel Embiid had 33 points and 10 rebounds in his return after missing three games with the flu,	Philadelphia
2802927188	2022-11-08 10:14:55+00:00	b'\xf0\x9f\x94\xb8Carrots for diabetics\nCarrots with their antioxidants&phytochemicals can keep ur blood sug	Mumbai, India
72299713	2022-11-08 10:14:54+00:00	b'RT @Maybell_N: Ah bueno... 78 a\xc3\xb1os, pas\xc3\xb3 hasta un c\xc3\xa1ncer de garganta y quieren que se v	Caracas, Venezuela
2983893126	2022-11-08 10:14:54+00:00	b'Flu'	Singapore
233192110	2022-11-08 10:14:54+00:00	b'@olamzzzz @funshodarq1 Lol...I have made my point...some people believe the earth is flat...and covid & car	London, England
1101633542163959808	2022-11-08 10:14:54+00:00	b'RT @Annette1Hardy: Well done, @JustStop_Oil. Our consultant surgeon friend is stuck on the M25 right now and unable to get to his clinic wh\xe2\x8	
1158934210146635777	2022-11-08 10:14:54+00:00	b'Atualizando: o outro ta com c\xc3\xa2ncer'	
2836533648	2022-11-08 10:14:53+00:00	b'RT @mansukhmandviya: \xe0\xa4\xb8\xe0\xa4\xb8\xe0\xa5\x8d\xe0\xa4\xa4\xe0\xa5\x80 \xe0\xa4\xa6\xe0\xa4\	हिंदुस्तान

Preprocessing Code

Indian Institute of Information Technology Kottayam

In [27]:
```
stop_words_file = 'SmartStoplist.txt'

stop_words = []

with open(stop_words_file, "r") as f:
    for line in f:
        stop_words.extend(line.split())

stop_words = stop_words
```

In [28]:
```
def preprocess(raw_text):

    #regular expression keeping only letters
    letters_only_text = re.sub("[^a-zA-Z]", " ", raw_text)

    # convert to lower case and split into words -> convert string into list ( 'hello world' -> ['hello', 'world'])
    words = letters_only_text.lower().split()

    cleaned_words = []
    lemmatizer = PorterStemmer() #plug in here any other stemmer or lemmatiser you want to try out

    # remove stopwords
    for word in words:
        if word not in stop_words:
            cleaned_words.append(word)

    # stemm or lemmatise words
    stemmed_words = []
    for word in cleaned_words:
        word = lemmatizer.stem(word)   #dont forget to change stem to lemmatize if you are using a lemmatizer
        stemmed_words.append(word)

    # converting list back to string
    return " ".join(stemmed_words)
```

Processed Data

```
df['split_tweets'] = df['prep'].str.split(' ')
```

```
df.head()
```

	id	Time	Tweet	Location	split	tags	split_tweets	prep
0	312322346	2022-11-08 10:14:36+00:00	b"Let's be honest; day two of a phase 1 cancer...	London	['b"Let\'s', 'be', 'honest;', 'day', 'two', 'o...	[cancer, thyroidcancer, thyroid]	[honest, day, phase, cancer, trial, look, good...	honest day phase cancer trial look good cancer...
1	42296887	2022-11-08 10:15:07+00:00	b'Pembrolizumab: English Drugs Body Recommends...	Florida	["b'Pembrolizumab:", 'English', 'Drugs', 'Body...	[]	[pembrolizumab, english, drug, bodi, recommend...	pembrolizumab english drug bodi recommend life...
2	1356818881063473157	2022-11-08 10:15:05+00:00	b"RT @nettybgoode: Well, @ServicesGovAU are at...	NaN	['b"RT', '@nettybgoode:', 'Well,', '@ServicesG...	[]	[rt, nettybgood, servicesgovau, day, mum, fini...	rt nettybgood servicesgovau day mum finish rad...
3	1570261324554293249	2022-11-08 10:15:04+00:00	b'@AlHosnApp where we can take session flu vac...	NaN	["b'@AlHosnApp", 'where', 'we', 'can', 'take',...	[]	[alhosnapp, session, flu, vaccin, guid]	alhosnapp session flu vaccin guid
4	4836684705	2022-11-08 10:15:03+00:00	b"RT @DFisman: Again, your immune system didn'...	Scotland, United Kingdom	['b"RT', '@DFisman:', 'Again,', 'your', 'immun...	[]	[rt, dfisman, immun, system, didn, flabbi, flu...	rt dfisman immun system didn flabbi flu season...

Indian Institute of Information Technology Kottayam

Experiment

Various Methods of sentiment analysis

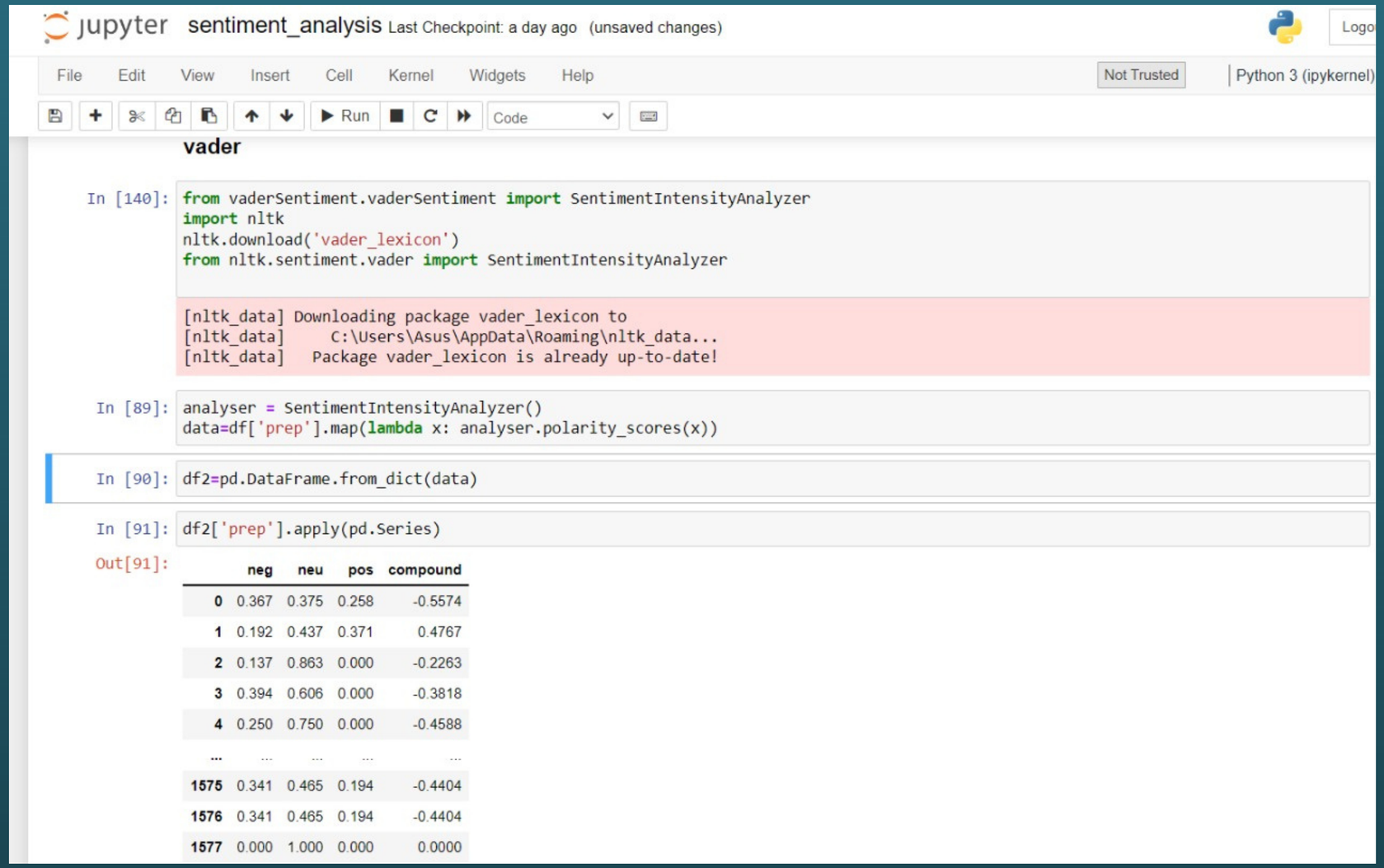
Using Text blob

```
df['split_tweets'] = df['prep'].str.split(' ')
```

```
df.head()
```

	id	Time	Tweet	Location	split	tags	split_tweets	prep
0	312322346	2022-11-08 10:14:36+00:00	b"Let's be honest; day two of a phase 1 cancer...	London	['b"Let\'s', 'be', 'honest;', 'day', 'two', 'o...	[cancer, thyroidcancer, thyroid]	[honest, day, phase, cancer, trial, look, good...	honest day phase cancer trial look good cancer...
1	42296887	2022-11-08 10:15:07+00:00	b'Pembrolizumab: English Drugs Body Recommends...	Florida	["b'Pembrolizumab:", 'English', 'Drugs', 'Body...	[]	[pembrolizumab, english, drug, bodi, recommend...	pembrolizumab english drug bodi recommend life...
2	1356818881063473157	2022-11-08 10:15:05+00:00	b"RT @nettybgoode: Well, @ServicesGovAU are at...	NaN	['b"RT', '@nettybgoode:', 'Well,', '@ServicesG...	[]	[rt, nettybgood, servicesgovau, day, mum, fini...	rt nettybgood servicesgovau day mum finish rad...
3	1570261324554293249	2022-11-08 10:15:04+00:00	b'@AlHosnApp where we can take session flu vac...	NaN	["b'@AlHosnApp", 'where', 'we', 'can', 'take',...	[]	[alhosnapp, session, flu, vaccin, guid]	alhosnapp session flu vaccin guid
4	4836684705	2022-11-08 10:15:03+00:00	b"RT @DFisman: Again, your immune system didn'...	Scotland, United Kingdom	['b"RT', '@DFisman:', 'Again,', 'your', 'immun...	[]	[rt, dfisman, immun, system, didn, flabbi, flu...	rt dfisman immun system didn flabbi flu season...

Using VADER



	neg	neu	pos	compound
0	0.367	0.375	0.258	-0.5574
1	0.192	0.437	0.371	0.4767
2	0.137	0.863	0.000	-0.2263
3	0.394	0.606	0.000	-0.3818
4	0.250	0.750	0.000	-0.4588
...	...	...	...	...
1575	0.341	0.465	0.194	-0.4404
1576	0.341	0.465	0.194	-0.4404
1577	0.000	1.000	0.000	0.0000

Using Normalization

jupyter Using Positive and Negative Word Count (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

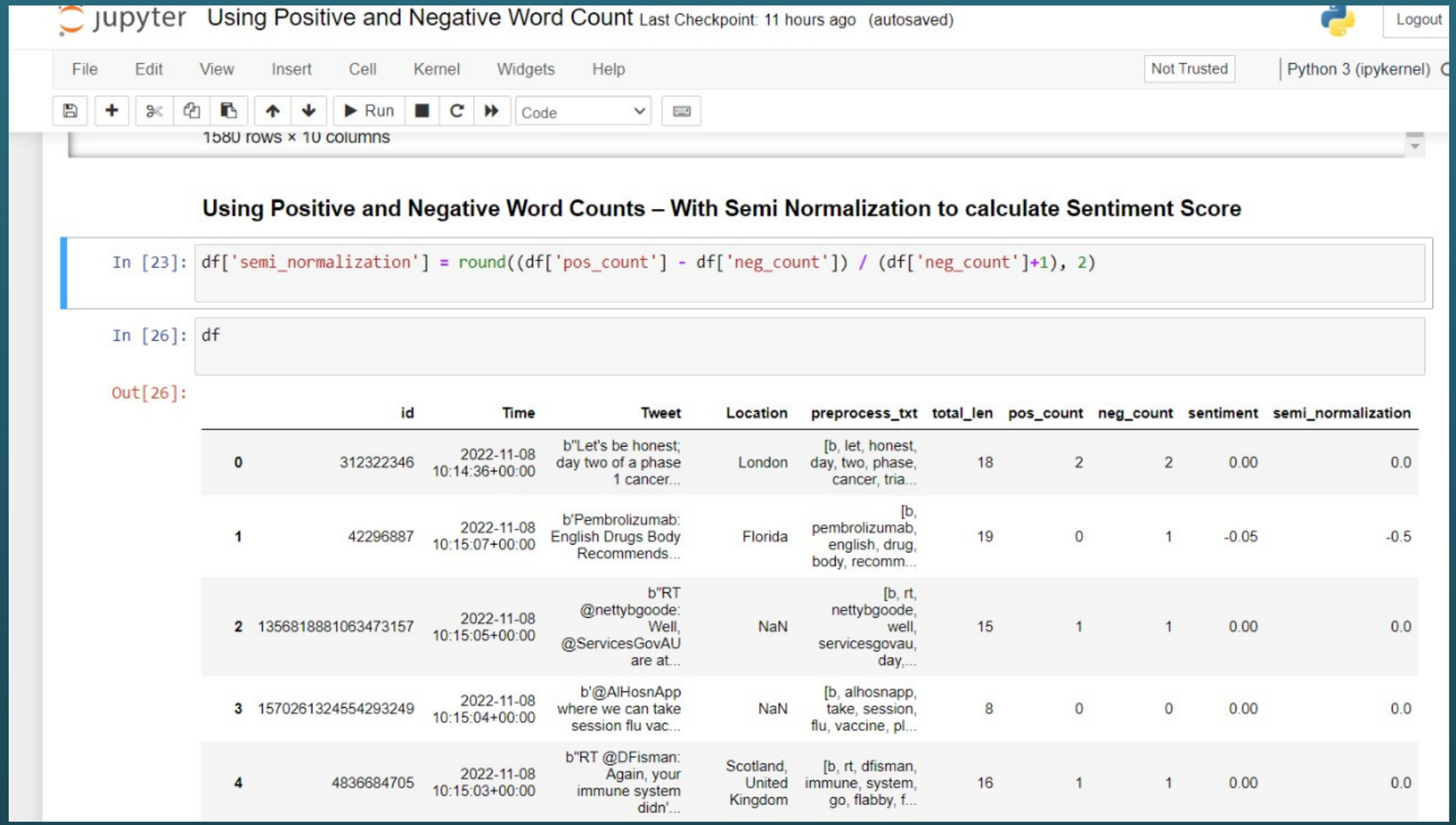
Run Code

In [21]:
```
df['sentiment'] = round((df['pos_count'] - df['neg_count']) / df['total_len'], 2
```

In [22]:
```
df.head()
```

id	Time	Tweet	Location	preprocess_txt	total_len	pos_count	neg_count	sentiment
312322346	2022-11-08 10:14:36+00:00	b"Let's be honest; day two of a phase 1 cancer...	London	[b, let, honest, day, two, phase, cancer, tria...	18	2	2	0.00
42296887	2022-11-08 10:15:07+00:00	b'Pembrolizumab: English Drugs Body Recommends...	Florida	[b, pembrolizumab, english, drug, body, recomm...	19	0	1	-0.05
1063473157	2022-11-08 10:15:05+00:00	b"RT @nettybgoode: Well, @ServicesGovAU are at...	NaN	[b, rt, nettybgoode, well, servicesgovau, day,...	15	1	1	0.00
4554293249	2022-11-08 10:15:04+00:00	b'@AlHosnApp where we can take session flu vac...	NaN	[b, alhosnapp, take, session, flu, vaccine, pl...	8	0	0	0.00

Using Semi Normalization



jupyter Using Positive and Negative Word Count Last Checkpoint: 11 hours ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

1580 rows × 10 columns

Using Positive and Negative Word Counts – With Semi Normalization to calculate Sentiment Score

In [23]:
```
df['semi_normalization'] = round((df['pos_count'] - df['neg_count']) / (df['neg_count']+1), 2)
```

In [26]:
```
df
```

Out[26]:

	id	Time	Tweet	Location	preprocess_txt	total_len	pos_count	neg_count	sentiment	semi_normalization
0	312322346	2022-11-08 10:14:36+00:00	b"Let's be honest; day two of a phase 1 cancer...	London	[b, let, honest, day, two, phase, cancer, tria...	18	2	2	0.00	0.0
1	42296887	2022-11-08 10:15:07+00:00	b'Pembrolizumab: English Drugs Body Recommends...	Florida	[b, pembrolizumab, english, drug, body, recomm...	19	0	1	-0.05	-0.5
2	1356818881063473157	2022-11-08 10:15:05+00:00	b"RT @nettybgoode: Well, @ServicesGovAU are at...	NaN	[b, rt, nettybgoode, well, servicesgovau, day,...	15	1	1	0.00	0.0
3	1570261324554293249	2022-11-08 10:15:04+00:00	b'@AlHosnApp where we can take session flu vac...	NaN	[b, alhosnapp, take, session, flu, vaccine, pl...	8	0	0	0.00	0.0
4	4836684705	2022-11-08 10:15:03+00:00	b"RT @DFisman: Again, your immune system didn'...	Scotland, United Kingdom	[b, rt, dfisman, immune, system, go, flabby, f...	16	1	1	0.00	0.0

Indian Institute of Information Technology Kottayam

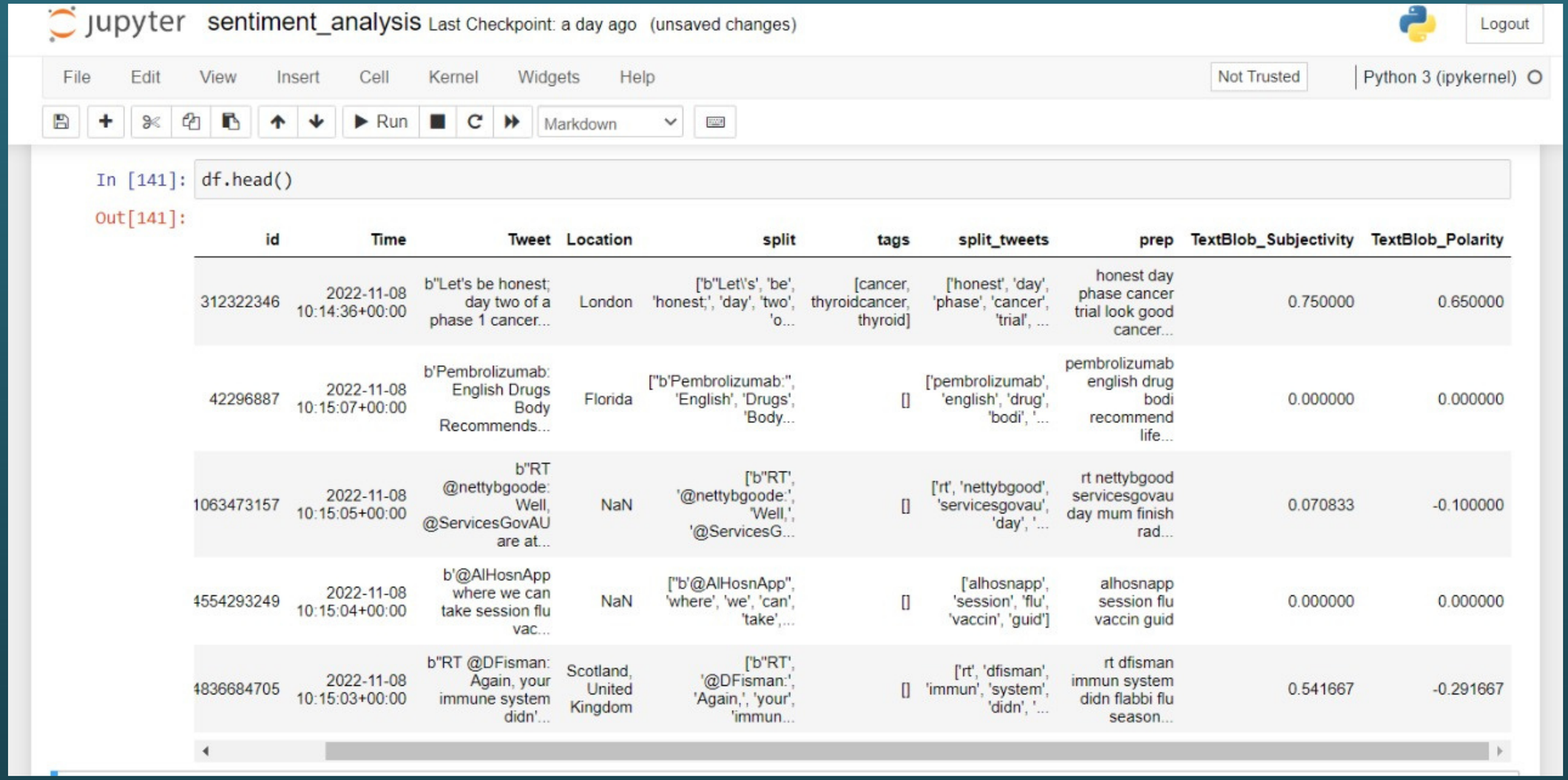
Demonstration

Indian Institute of Information Technology Kottayam

Result

Results of various sentiment analysis methods

Using Text blob


jupyter sentiment_analysis Last Checkpoint: a day ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

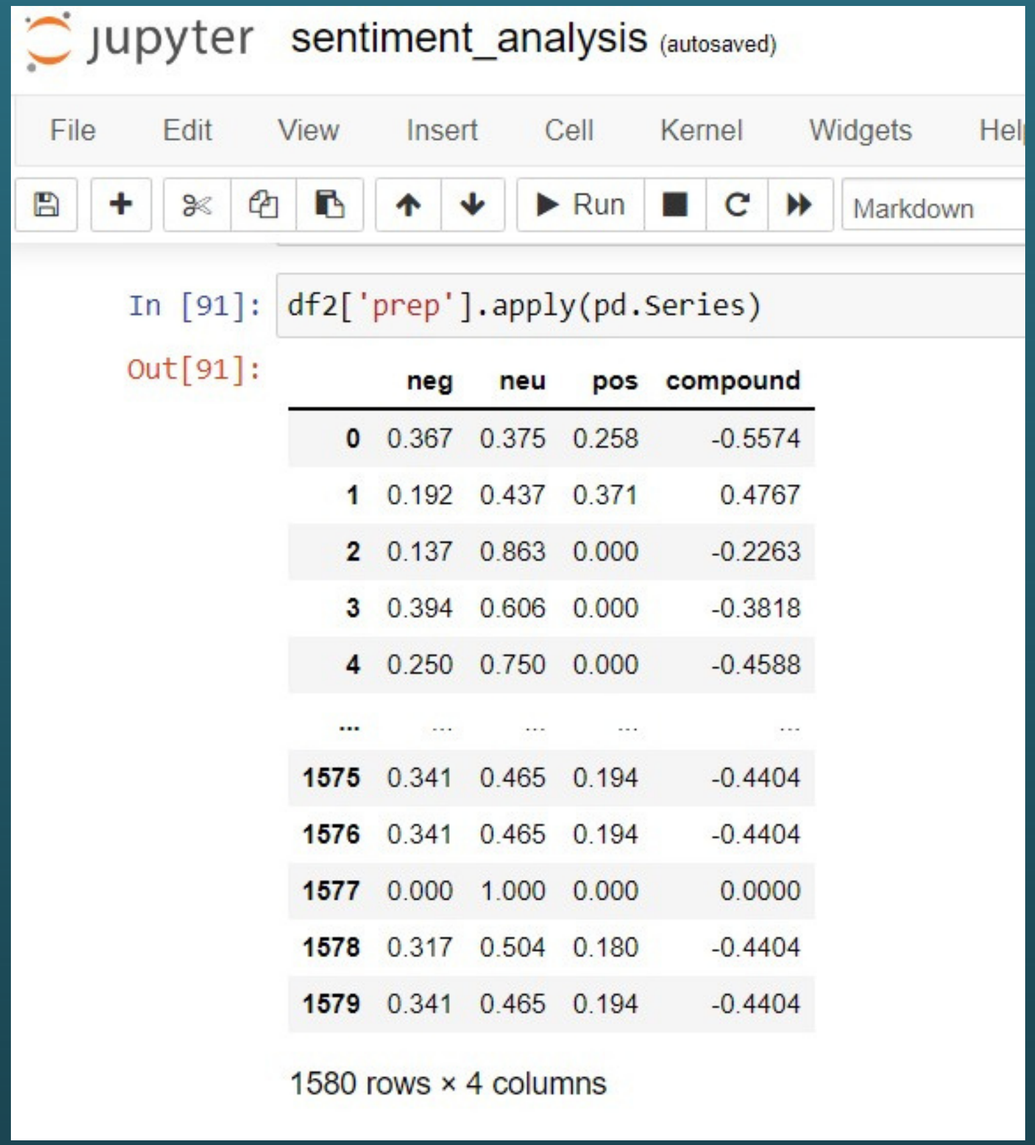
Run Markdown

In [141]:
```
df.head()
```

Out[141]:

	id	Time	Tweet	Location	split	tags	split_tweets	prep	TextBlob_Subjectivity	TextBlob_Polarity
	312322346	2022-11-08 10:14:36+00:00	b"Let's be honest; day two of a phase 1 cancer...	London	['b"Let\'s', 'be', 'honest;', 'day', 'two', 'o...	[cancer, thyroidcancer, thyroid]	['honest', 'day', 'phase', 'cancer', 'trial', ...	honest day phase cancer trial look good cancer...	0.750000	0.650000
	42296887	2022-11-08 10:15:07+00:00	b'Pembrolizumab: English Drugs Body Recommends...	Florida	["b'Pembrolizumab:", 'English', 'Drugs', 'Body...	[]	['pembrolizumab', 'english', 'drug', 'bodi', '...	pembrolizumab english drug bodi recommend life...	0.000000	0.000000
	1063473157	2022-11-08 10:15:05+00:00	b"RT @nettybgoode: Well, @ServicesGovAU are at...	NaN	['b"RT', '@nettybgoode:', 'Well,', '@ServicesG...	[]	['rt', 'nettybgood', 'servicesgovau', 'day', '...	rt nettybgood servicesgovau day mum finish rad...	0.070833	-0.100000
	4554293249	2022-11-08 10:15:04+00:00	b'@AlHosnApp where we can take session flu vac...	NaN	["b'@AlHosnApp", 'where', 'we', 'can', 'take',...	[]	['alhosnapp', 'session', 'flu', 'vaccin', 'guid']	alhosnapp session flu vaccin guid	0.000000	0.000000
	4836684705	2022-11-08 10:15:03+00:00	b"RT @DFisman: Again, your immune system didn'...	Scotland, United Kingdom	['b"RT', '@DFisman:', 'Again,', 'your', 'immun...	[]	['rt', 'dfisman', 'immun', 'system', 'didn', '...	rt dfisman immun system didn flabbi flu season...	0.541667	-0.291667

Using VADER



	neg	neu	pos	compound
0	0.367	0.375	0.258	-0.5574
1	0.192	0.437	0.371	0.4767
2	0.137	0.863	0.000	-0.2263
3	0.394	0.606	0.000	-0.3818
4	0.250	0.750	0.000	-0.4588
...	...	...	...	...
1575	0.341	0.465	0.194	-0.4404
1576	0.341	0.465	0.194	-0.4404
1577	0.000	1.000	0.000	0.0000
1578	0.317	0.504	0.180	-0.4404
1579	0.341	0.465	0.194	-0.4404

1580 rows × 4 columns

Using Normalization

```
In [22]: df.head()
```

Out[22]:

	id	Time	Tweet	Location	preprocess_txt	total_len	pos_count	neg_count	sentiment
0	312322346	2022-11-08 10:14:36+00:00	b"Let's be honest; day two of a phase 1 cancer...	London	[b, let, honest, day, two, phase, cancer, tria...	18	2	2	0.00
1	42296887	2022-11-08 10:15:07+00:00	b'Pembrolizumab: English Drugs Body Recommends...	Florida	[b, pembrolizumab, english, drug, body, recomm...	19	0	1	-0.05
2	1356818881063473157	2022-11-08 10:15:05+00:00	b"RT @nettybgoode: Well, @ServicesGovAU are at...	NaN	[b, rt, nettybgoode, well, servicesgovau, day,...	15	1	1	0.00
3	1570261324554293249	2022-11-08 10:15:04+00:00	b'@AlHosnApp where we can take session flu vac...	NaN	[b, alhosnapp, take, session, flu, vaccine, pl...	8	0	0	0.00
4	4836684705	2022-11-08 10:15:03+00:00	b"RT @DFisman: Again, your immune system didn'...	Scotland, United Kingdom	[b, rt, dfisman, immune, system, go, flabby, f...	16	1	1	0.00

Using Semi Normalization

```
In [24]: df.head()
```

Out[24]:

	id	Time	Tweet	Location	preprocess_txt	total_len	pos_count	neg_count	sentiment	semi_normalization
0	312322346	2022-11-08 10:14:36+00:00	b"Let's be honest; day two of a phase 1 cancer...	London	[b, let, honest, day, two, phase, cancer, tria...	18	2	2	0.00	0.0
1	42296887	2022-11-08 10:15:07+00:00	b'Pembrolizumab: English Drugs Body Recommends...	Florida	[b, pembrolizumab, english, drug, body, recomm...	19	0	1	-0.05	-0.5
2	1356818881063473157	2022-11-08 10:15:05+00:00	b"RT @nettybgoode: Well, @ServicesGovAU are at...	NaN	[b, rt, nettybgoode, well, servicesgovau, day,...	15	1	1	0.00	0.0
3	1570261324554293249	2022-11-08 10:15:04+00:00	b'@AlHosnApp where we can take session flu vac...	NaN	[b, alhosnapp, take, session, flu, vaccine, pl...	8	0	0	0.00	0.0
4	4836684705	2022-11-08 10:15:03+00:00	b"RT @DFisman: Again, your immune system didn'...	Scotland, United Kingdom	[b, rt, dfisman, immune, system, go, flabby, f...	16	1	1	0.00	0.0

Indian Institute of Information Technology Kottayam

Conclusion

- Until now four different methods were used to calculate the sentiment score of a given text. We would like to explore more methods for the same.
- Validating accuracy of the Sentiment score calculated by following the above mentioned methods can be done by having a labeled (annotated) training data set. During model construction we usually use an annotated data set. The accuracy can be checked by comparing annotated test records.
- After validating the accuracy for all the methods, we will choose the one with the highest accuracy for our data set.
- After calculating the sentiment score of a tweet it will then be divided into positive or negative categories based on a pre defined threshold value of sentiment score.
- The next step is feature summarization based on which activity recommendation will be done to the user.

Indian Institute of Information Technology Kottayam

References

[1]Abbi Nizar Muhammad, Saiful Bukhori, Priza Pandunata. Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes – Support Vector Machine (NBSVM) Classifier.
[2]Sam Liu, Brian Chen, Alex Kuo. Monitoring Physical Activity Levels Using Twitter Data: Infodemiology Study.
[3]Sunil Gohil et. al. Sentiment analysis of health care tweets: Review of the methods used, 2015.
[4] Marjolijn L. Antheunis a et. al. Patients' and health professionals' use of social media in health care: Motives, barriers and expectations, 2017.
[5]C.J. Hutto et. al. Vader: A parsimonious rule-based model for sentiment analysis of social media, 2016.
[6]Nazia Zaman et. al. Semantics-enhanced recommendation system for social healthcare, 2019.

Indian Institute of Information Technology Kottayam

Thank You