

**TWITTER – USERS’ CONTENT  
RECOMMENDATION SYSTEM USING TWEETS**  
**A PROJECT REPORT**

*Submitted by*

**SHREENIDHI S                      2017103620**  
**PAVITHRA KARTHY            2017103059**  
**SUPRADEEPA VELLA        2017103628**

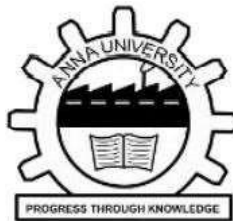
*in partial fulfilment for the award of the degree*

*of*

**BACHELOR OF ENGINEERING**

**IN**

**COMPUTER SCIENCE AND ENGINEERING**



**COLLEGE OF ENGINEERING, GUINDY**

**ANNA UNIVERSITY: CHENNAI 600 025**

**NOVEMBER 2020**

## **TABLE OF CONTENTS**

<b>CHAPTER NO.</b>	<b>TITLE</b>	<b>PAGE NO.</b>
<b>1.</b>	<b>PROBLEM STATEMENT</b>	<b>1</b>
<b>2.</b>	<b>ABSTRACT</b>	<b>1</b>
<b>3.</b>	<b>BASE PAPER</b>	<b>2</b>
<b>4.</b>	<b>ABOUT CODE</b>	<b>2</b>
<b>5.</b>	<b>MODIFICATIONS DONE</b>	<b>2</b>
<b>6.</b>	<b>DATASET</b>	<b>2</b>
<b>7.</b>	<b>METHODOLOGY</b>	<b>3</b>
<b>8.</b>	<b>CONFIGURATION PARAMETERS</b>	<b>6</b>
<b>9.</b>	<b>RESULTS</b>	<b>7</b>
<b>10.</b>	<b>EVALUATION</b>	<b>8</b>

# 1. PROBLEM STATEMENT

The aim of this project is to build a recommendation system that can help an organization to target their customers based on their specific interests. The main objectives of our recommendation system include:

- Sentimental analysis of individual users
- Visualization of individual user interests
- Grouping of users with same interests
- Comparison of users with same interests

# 2. ABSTRACT

With the advent of the internet into our everyday lives, online social networks such as Facebook and Twitter have taken up a major role in networking, information deployment and entertainment. As of 2017, Twitter's outreach is over 317M monthly active users generating more than 320M tweets everyday, thus making it one of the fastest information deployment mediums of this era. In order to aid data distribution without causing a glut of information to the users, we develop a recommender system focusing on a vital aspect of most preferred interests of the individual users. The information collected from the most recent tweets of a user is used to find other users whose recent tweets contain similar information, and grouping of users with common interests are also done. By making use of the continuous and real time updating of data on social networks, we develop a method to ensure our training sets consist of relevant information for classification, thus preserving accuracy while reducing training set sizes for probabilistic learning models. We use two methodologies to detect tweets of common topics, namely a Lemmatization and a Cosine similarity Classifier and further compare their complexity and accuracy.

The information collected depicts the percentage of the most tweeted and less tweeted topics users are interested in and then grouping other users with common ideas and interests in common categories and the individual interests of the users are shown in the pie chart.

The aim of this project is to build a recommendation system that can help an organization to target their customers based on their specific interests. This can help the organization in forming effective marketing strategies to attract customers. Users interests are gathered from their twitter account in the form of tweets, favourites etc. For collecting data, we used tweepy API to extract tweets, favourites etc. Then we classified the users interests into relevant categories with an accuracy of about 85% by using NLP algorithms. We also performed sentiment analysis of tweets, retweets and were able to achieve an accuracy of 90%. Individual user interests are displayed using pie chart built using matplotlib whereas overall user interests are displayed using bar graph built using bokeh.

### 3. BASE PAPER

#### 1) Short and Tweet: Experiments on Recommending Content from Information Streams

**Authors:** Jilin Chen

**LINK:** <https://hci.stanford.edu/publications/2010/zerozero88/zerozero88-chi2010.pdf>

#### 2) Twitter-User Recommender System using Tweets: A Content-based Approach

**Authors:** Nidhi R.H, Annappa B

**Link:** <https://ieeexplore.ieee.org/document/8272631>

### 4. ABOUT CODE

#### FULL CODE:

[https://github.com/snidhi99/TWITTER\\_SENTIMENT\\_ANALYSIS/blob/master/Recommendation\\_System\\_User\\_Interests\(1\).ipynb](https://github.com/snidhi99/TWITTER_SENTIMENT_ANALYSIS/blob/master/Recommendation_System_User_Interests(1).ipynb)

The code is not the same as the research paper.

#### Research Paper idea

- In the research paper, the users are categorized based on their popularity and the information collected from the most recent tweets are used to find other users whose tweets contain similar information.
- Naïve bayes classification is done
- So, it is a user recommender system based on tweets

#### Our Project idea

- In our project, we are categorizing the users based on their specific interests, i.e, based on the content that they tweet and are visualizing each users' interest through a pie chart.
- Ultimately, the most tweeted ideas correspond to their category interests.
- This is what we would recommend, by eventually grouping them under interested categories.
- Additionally, sentimental analysis followed by lemmatization and cosine similarity, is performed

### 5. MODIFICATIONS DONE

- Sentimental Analysis – Classification of tweets, retweets and favorite tweets as positive and negative
- Getting Topwords and categories – identifying highest frequency words and putting them under their respective categories
- Bag of words - All the unwanted words and articles are removed, making it easier for us to identify the important keywords
- Lemmatization - process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's lemma, from the categories list

- Cosine Similarity - where we will have to convert the word to its vector form, and we will also have to convert the word in the categories list to its vector form

## 6. DATASET

### (1) User\_Set\_1.csv

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	twitter handle																		
2	@BarackObama																		
3	@katyperry																		
4	@Justinbieber																		
5	@rihanna																		
6	@taylorswift13																		
7	@Cristiano																		
8	@ladygaga																		
9	@TheEllenShow																		
10	@YouTube																		
11	@ArianaGrande																		
12	@realDonaldTrump																		
13	@jimmyfallon																		
14	@KimKardashian																		
15	@selenagomez																		
16	@Twitter																		
17	@britneyspears																		
18	@cnnbrk																		
19	@narendramodi																		
20	@shakira																		
21	@jimmyfallon																		
22	@BillGates																		
23	@ninemajr																		
24	@mattlauer																		

#### DESCRIPTION:

- Twitter handles (User names) – Provides a list of all popular twitter users

### (2) Categories\_List.csv

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Sports	Politics	Finance	Travel	Informatic	History	Medical	Story	React	Food	Entertain	Shopping	Education						
2	athletics	governme	funds	trip	ict	chronicle	field of st	Hawthorn	feed	treat	mail	teaching							
3	sportsmar	political	sc	cash in	journey	computer:story	Pediatric	Mann	grub	entertain	shopper	pedagogy							
4	sportswor	political	re	Monetary	jaunt	informatic:account	Islamic	La	boll	chow	cabaret	grocery	instruction						
5	fun	political	spec	unary	go	icts	annals	child	psycl	Hemingw	victual	amuse	supermar	training					
6	play	political	treasury	move	computer	lore	iatro	Hesse	provender	divertisser	retail	department of education							
7	disport	politician	banking	locomotio	computing	record	iatromath	Tolstoy	repast	masque	store	breeding							
8	lark	politicians	financial	change of	data-pro	heritage	group	pra	readable	eat	nightlife	stores	education	department					
9	boast	polity	economic	locomote	Edp	ever	medicine	fiction	tucker	perform	buying	Educational	Activity						
10	nonreside	politically	loans	travelers	technology	historical	hippocraft	read	cook	Scout	purchases	curriculum							
11	cavort	affairs	treasurer	trips	computer	tradition	leechcraft	west	tuck	interlude	purchasin	schooling							
12	frilick	Politik	lending	journeys	computer	historiography	iatry	plot	fast	nightclub	business	literacy							
13	mutation	intrigues	accountin	flights	data	historic	pathology	plotted	Digest	showbusi	commerce	curricula							
14	frilic	policy	fund	traveler	electronic	historian	acology	plotting	eatable	vaudeville	buy	schools							
15	skylark	policies	loan	itinerary	software	chronolog	zoopathol	roman	fle	hunger	show	errand	educators						
16	romp	motivation	financier	cruise	telematics	legacy	dietetist	copy	nourishm	junket	consumer	teachers							
17	gambol	policies	investmen	fly	systems	records	Scientific	cartoonist	slaughter	host	spending	Academic							
18	mutant	policymak	underwrit	vacation	Cti	historical	oncology	literature	aliment	carnival	plaza	tuition							
19	feature	policy-mal	capital	commutin	infotech	anteceder	medical	Malraux	scoff	party	purchase	school							
20	lark	about	policymak	tax	sightseein	Arda	past	iatrochem	Saroyan	masticate	revue	arcade	educator						
21	run	aroun	Politica	credit	hitchhike	ntis	time	physiatri	wells	Fed	entertaini	trading	diploma						
22	frisk	politique	economic	destinatio	tick	evolution	tocology	novel	forage	karaoke	sourcing	teacher							
23	athletic	politico	money	commute	bioinform	experieno	physiologi	ana	dish	escapism	marketing	students							
24	athlatic	academic	cash	transpanti	communicati	doctor	of	credit	nurture	extravagan	programe								

#### DESCRIPTION:

- The various categories are the independent columns
- Each column has a list of all common verb forms
- Cosine similarity will be checked if the word matches with any of the words of the list

## 7. METHODOLOGY

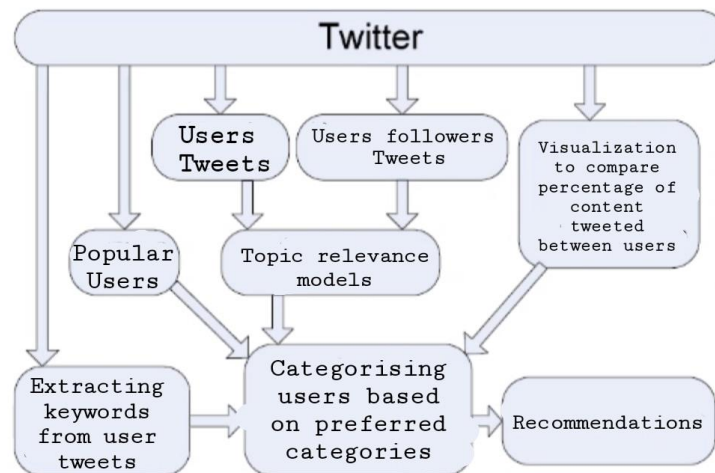


Figure 1. Conceptual Model of the Whole Recommender

### (1) API AUTHENTICATION

- Generated data from twitter streams and then converted into a csv file
- Importing Data From Pre-Defined Datasets And Writing The Output Recommendation To A New Csv File

### (2) FUNCTION TO EXTRACT TWEETS, RETWEETS AND FAVOURITE TWEETS

- The API.retweets() method of the API class in Tweepy module is used to return a list of retweets of a tweet.
- Using the retweets() method with count parameter to only fetch a certain number of retweets.

### (3) FUNCTION TO PERFORM SENTIMENTAL ANALYSIS OF TWEETS

- The tweets are classified as positive/negative

### (4) TO GET TOPWORDS

- Implementation of a function that gets the highest frequency words from the unigram, bigram and trigrams list

### (5) FIND WHAT CATEGORY THE WORDS WILL BELONG TO

- Implementation of a function that helps to keep track of the number of words that belong to specific categories

### (6) CLASSIFY WORDS BASED ON CATEGORIES/PERCENTAGE OF OCCURANCE

- categoryWisePercent function that classifies words belonging to each category and calculates categoryWise percentage

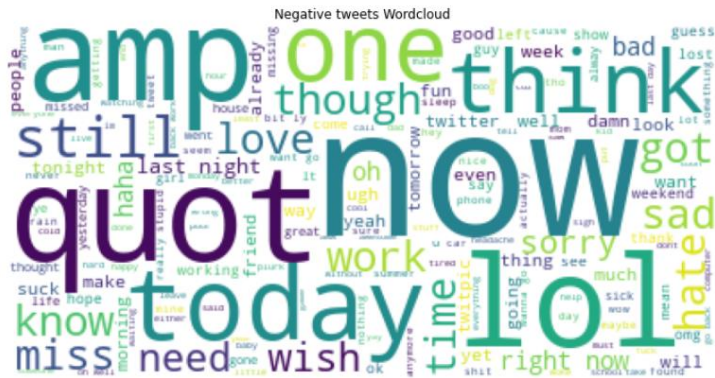
### (7) PLOT USERS' RECOMMENDATIONS AS PIE CHARTS



### (3) Negative Tweets (Content)

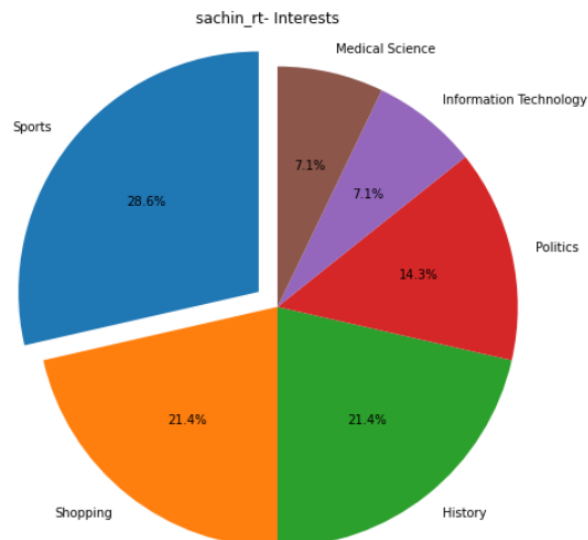
```
In [ ]: wordcloud = WordCloud(stopwords = STOPWORDS, background_color = "white", max_words = 1000).generate(negative_sentiments)
plt.figure(figsize = (12, 8))
plt.imshow(wordcloud)
plt.axis("off")
plt.title("Negative tweets Wordcloud")
```

```
Out[ ]: Text(0.5, 1.0, 'Negative tweets Wordcloud')
```



**Sample pie charts of different users and their category interests are depicted:**  
(1)

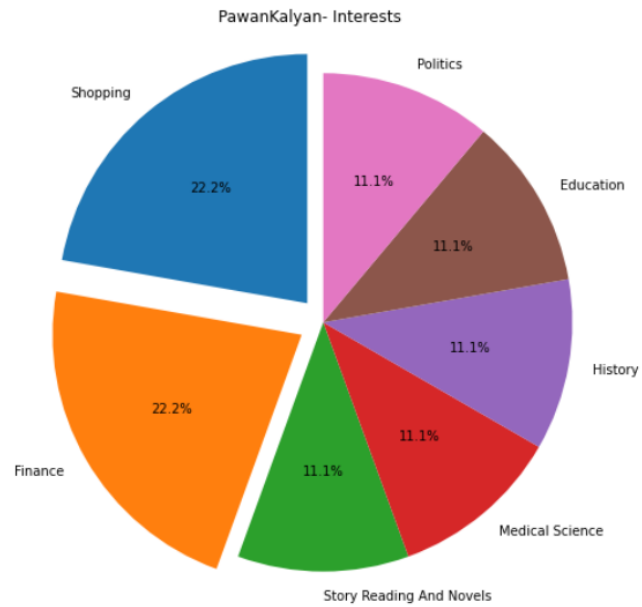
User Name : sachin_rt			
	Category	Words	Percentage
0	Sports	4	28
1	Shopping	3	21
2	History	3	21
3	Politics	2	14
4	Information Technology	1	7
5	Medical Science	1	7



(2)

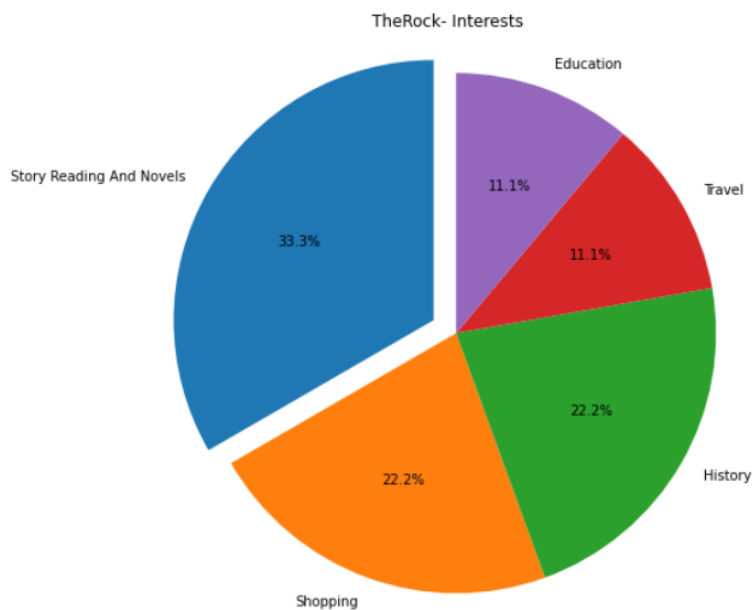


User Name : PawanKalyan			
	Category	Words	Percentage
0	Shopping	2	22
1	Finance	2	22
2	Story Reading And Novels	1	11
3	Medical Science	1	11
4	History	1	11
5	Education	1	11
6	Politics	1	11



(3)

User Name : TheRock			
	Category	Words	Percentage
0	Story Reading And Novels	3	33
1	Shopping	2	22
2	History	2	22
3	Travel	1	11
4	Education	1	11



Bar Chart to display individual users grouped into various categories. The number of users in each category and the total percentage of content used is tabulated.

The number of users' highest percentage of content tweeted, corresponding to each user's category interest will be recommended to the user.

Names	Categories	Number Of Users	Percentage	User
0	Sports	7	16	sachin_rt,virendersehwag,Cristiano,JustinTrudeau,realDonaldTrump,RaviShastriOf
c,KP24	Shopping	6	13	sachin_rt,virendersehwag,Cristiano,JustinTrudeau,PawanKalya
1	History	6	13	sachin_rt,virendersehwag,Cristiano,JustinTrudeau,realDonaldTrump
n,KP24	Travel	4	9	Cristiano,JustinTrudeau,realDonaldTrump
2	Story Reading And Novels	6	13	Cristiano,JustinTrudeau,lbjamesharden,realDonaldTrump,KP24,T
p,KP24	Finance	3	6	JustinTrudeau,PawanKalya
3	Food	2	4	JustinTrudeau,realDonaldTrump
4	Politics	4	9	JustinTrudeau,realDonaldTrump,KP24,PTTVOnli
heRock	Entertainment	3	6	realDonaldTrump,KP24,PTTVOnli
5	Education	1	2	realDonaldTrump
n,KP24	Medical Science	1	2	realDonaldTrump
6				
dTrump				
7				
neNews				
8				
neNews				
9				
dTrump				
10				
dTrump				

## 10. EVALUATION

### ALGORITHMS USED

- Content based filtering
- finding cosine correlation based similarities
- NLTK algorithms like stemming, lemitization