PROJECT DISCUSSION: TWITTER DATA SET ON HEALTH NEWS

JOSEPH TOCHUKWU EZEGWU

AWAL MUHAMMAD UMAR

NIMPA MBOUO RAYAUL

JOSHUA JUNIOR JOHN

DATA ANALYTICS AND DATA DRIVEN DECISION PROJECT 2021

OUTLINE

Introduction

Description of Data set

Data Manipulation

Data cleaning

Exploratory data Analysis

- Hashtag
- User-mention
- Word Occurrence

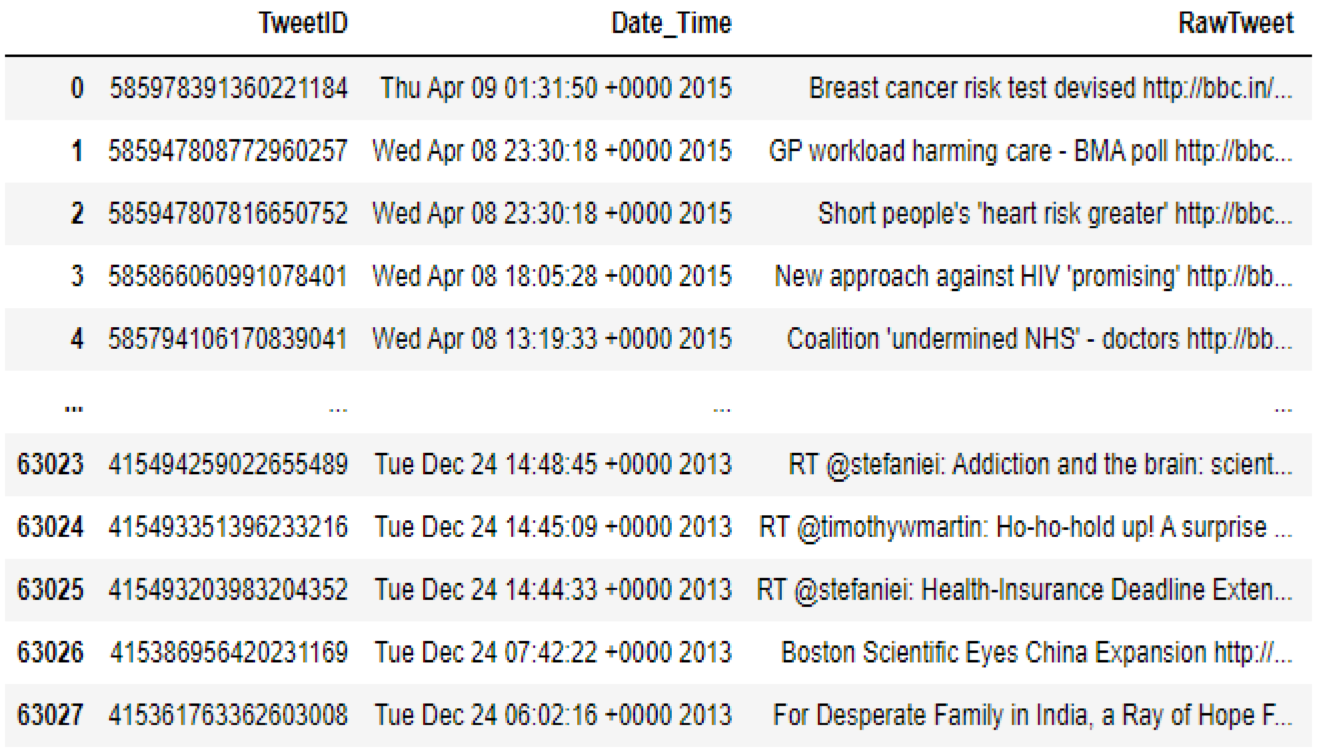
Discussion of Analysis by

1. Unsupervised learning Technique

Conclusion and Result

DESCRIPTION OF DATA SET

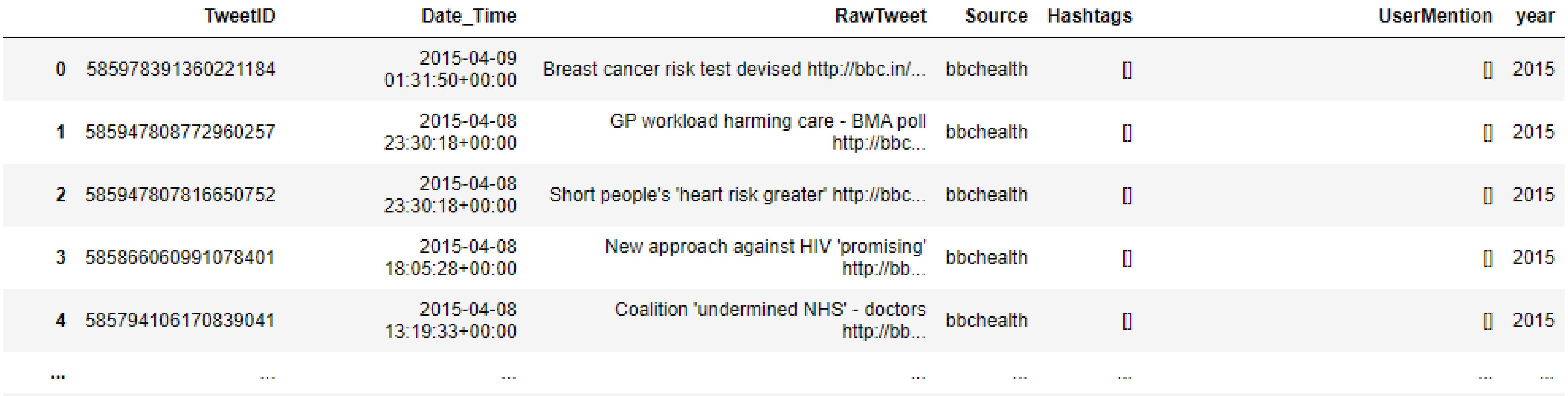
- Dataset contains health related tweets from 16 news agency.
- Dimension: 68000 rows by 3 Columns across 16 text files
- Data involves text input



	TweetID	Date_Time	RawTweet
0	585978391360221184	Thu Apr 09 01:31:50 +0000 2015	Breast cancer risk test devised http://bbc.in/...
1	585947808772960257	Wed Apr 08 23:30:18 +0000 2015	GP workload harming care - BMA poll http://bbc...
2	585947807816650752	Wed Apr 08 23:30:18 +0000 2015	Short people's 'heart risk greater' http://bbc...
3	585866060991078401	Wed Apr 08 18:05:28 +0000 2015	New approach against HIV 'promising' http://bb...
4	585794106170839041	Wed Apr 08 13:19:33 +0000 2015	Coalition 'undermined NHS' - doctors http://bb...
...	...	...	...
63023	415494259022655489	Tue Dec 24 14:48:45 +0000 2013	RT @stefaniei: Addiction and the brain: scient...
63024	415493351396233216	Tue Dec 24 14:45:09 +0000 2013	RT @timothywmartin: Ho-ho-hold up! A surprise ...
63025	415493203983204352	Tue Dec 24 14:44:33 +0000 2013	RT @stefaniei: Health-Insurance Deadline Exten...
63026	415386956420231169	Tue Dec 24 07:42:22 +0000 2013	Boston Scientific Eyes China Expansion http://...
63027	415361763362603008	Tue Dec 24 06:02:16 +0000 2013	For Desperate Family in India, a Ray of Hope F...

DATA MANIPULATION

- Extracted Year from Date_Time column
- Extracted Hashtags, and mentions from Raw tweet
- Created a new column to indicate the news agency (Source)



	TweetID	Date_Time	RawTweet	Source	Hashtags	UserMention	year
0	585978391360221184	2015-04-09 01:31:50+00:00	Breast cancer risk test devised http://bbc.in/...	bbchealth	[]	[]	2015
1	585947808772960257	2015-04-08 23:30:18+00:00	GP workload harming care - BMA poll http://bbc...	bbchealth	[]	[]	2015
2	585947807816650752	2015-04-08 23:30:18+00:00	Short people's 'heart risk greater' http://bbc...	bbchealth	[]	[]	2015
3	585866060991078401	2015-04-08 18:05:28+00:00	New approach against HIV 'promising' http://bb...	bbchealth	[]	[]	2015
4	585794106170839041	2015-04-08 13:19:33+00:00	Coalition 'undermined NHS' - doctors http://bb...	bbchealth	[]	[]	2015
...	...	...	...	...	...	...	...

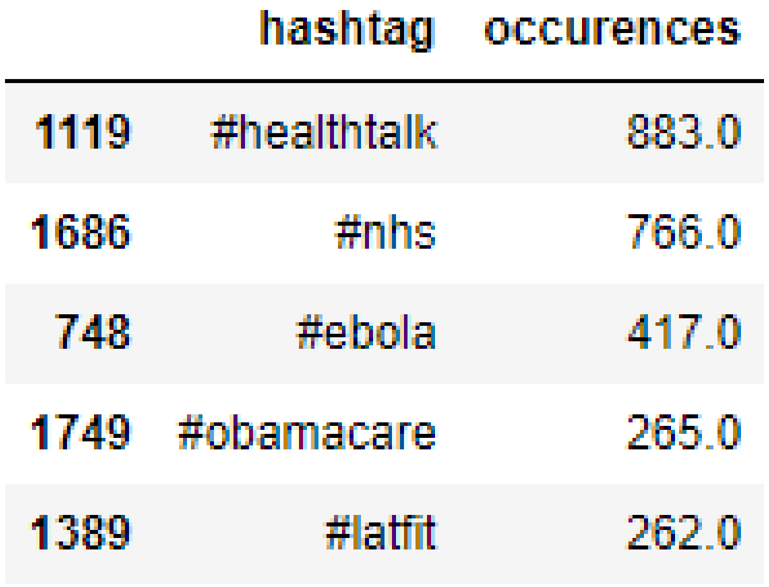
DATA CLEANING

Cleaned the raw tweets by:

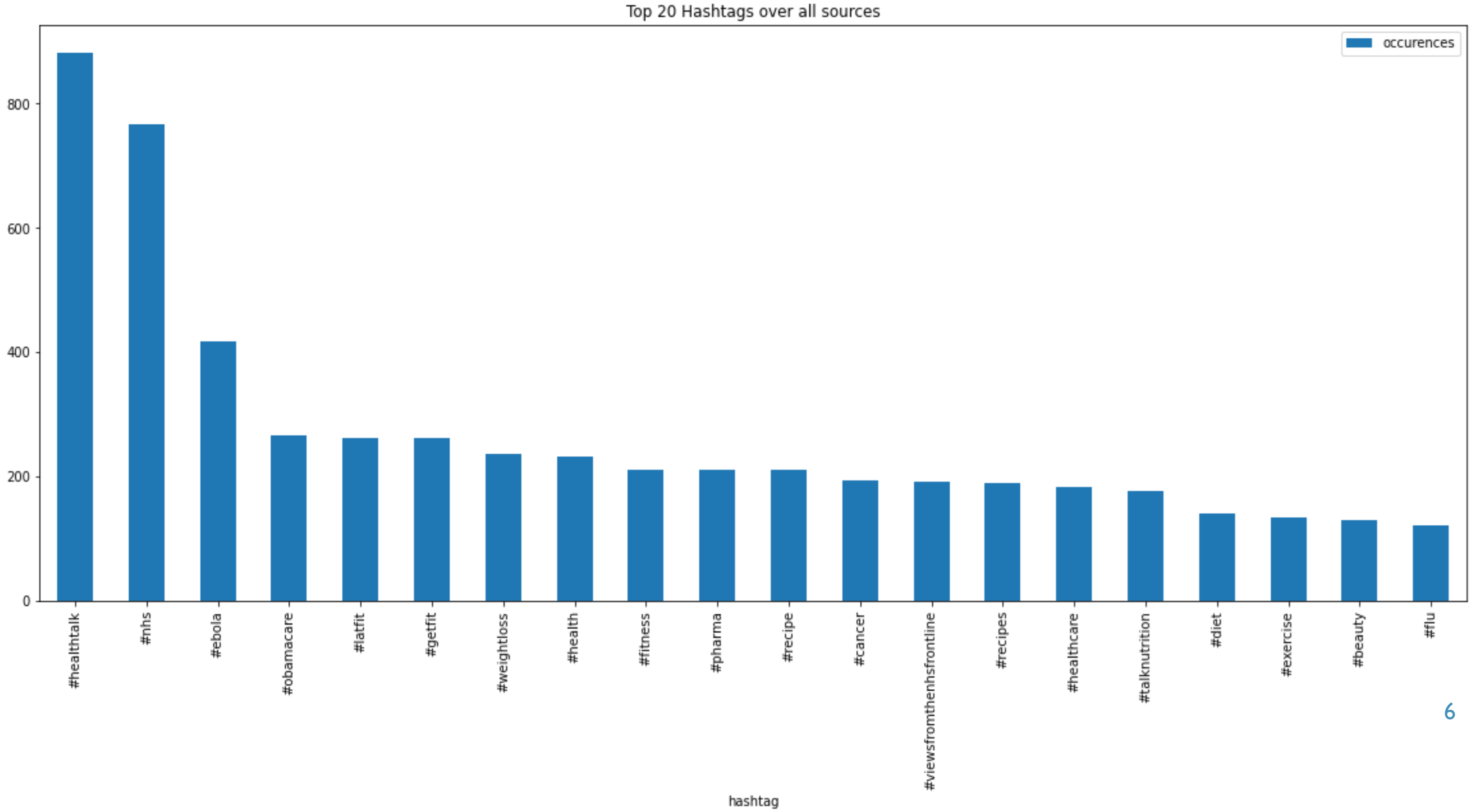
- Removing punctuation, English stopwords and tweeter stopwords (eg. rt, like, say etc.)
- Removing hashtags, links and mentions (eg. @username)
- Applying Stemming and lemmatization

EXPLORATORY DATA ANALYSIS (HASHTAGS EXPLORATORY ANALYSIS)

Top 20 trending Hashtags across all news agencies from 2011 to 2015.

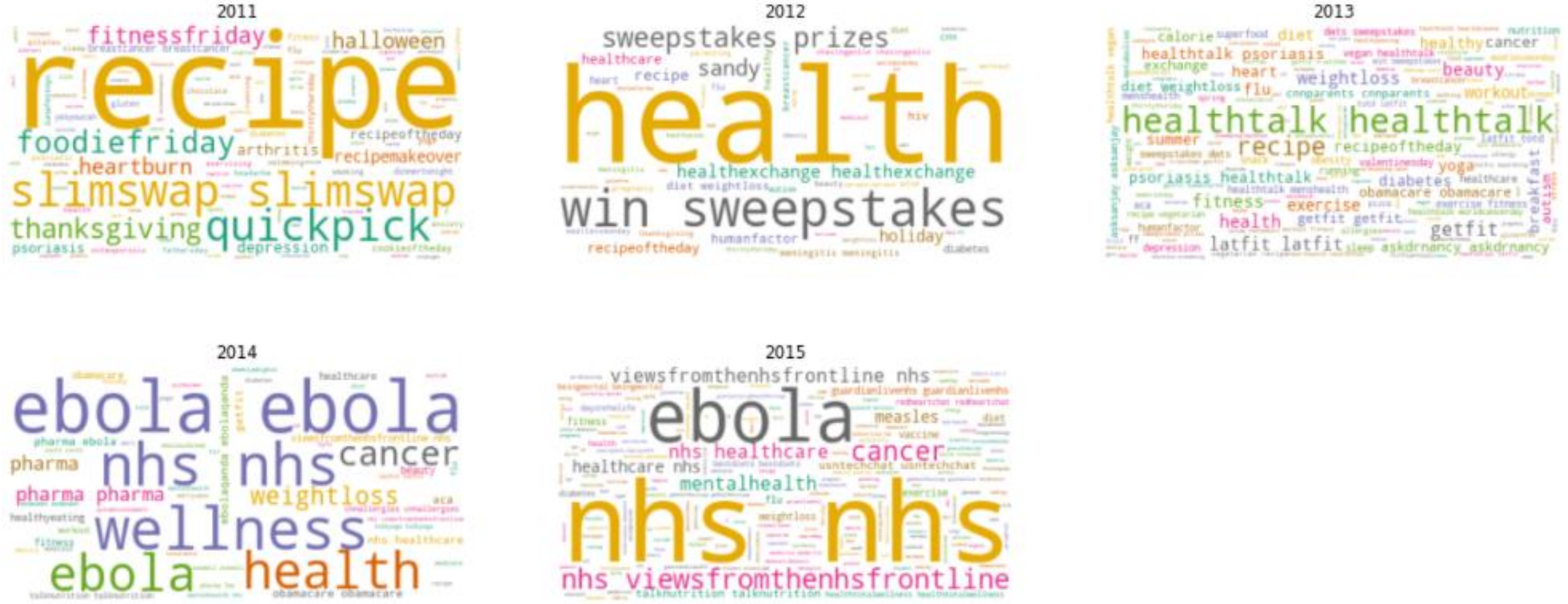


	hashtag	occurences
1119	#healthtalk	883.0
1686	#nhs	766.0
748	#ebola	417.0
1749	#obamacare	265.0
1389	#latfit	262.0

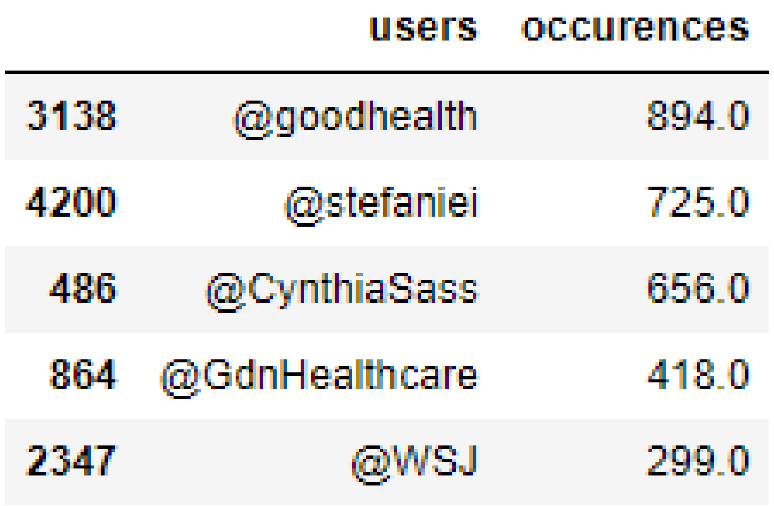


EXPLORATORY DATA ANALYSIS (HASHTAGS EXPLORATORY ANALYSIS)

Trending Hashtags By Year

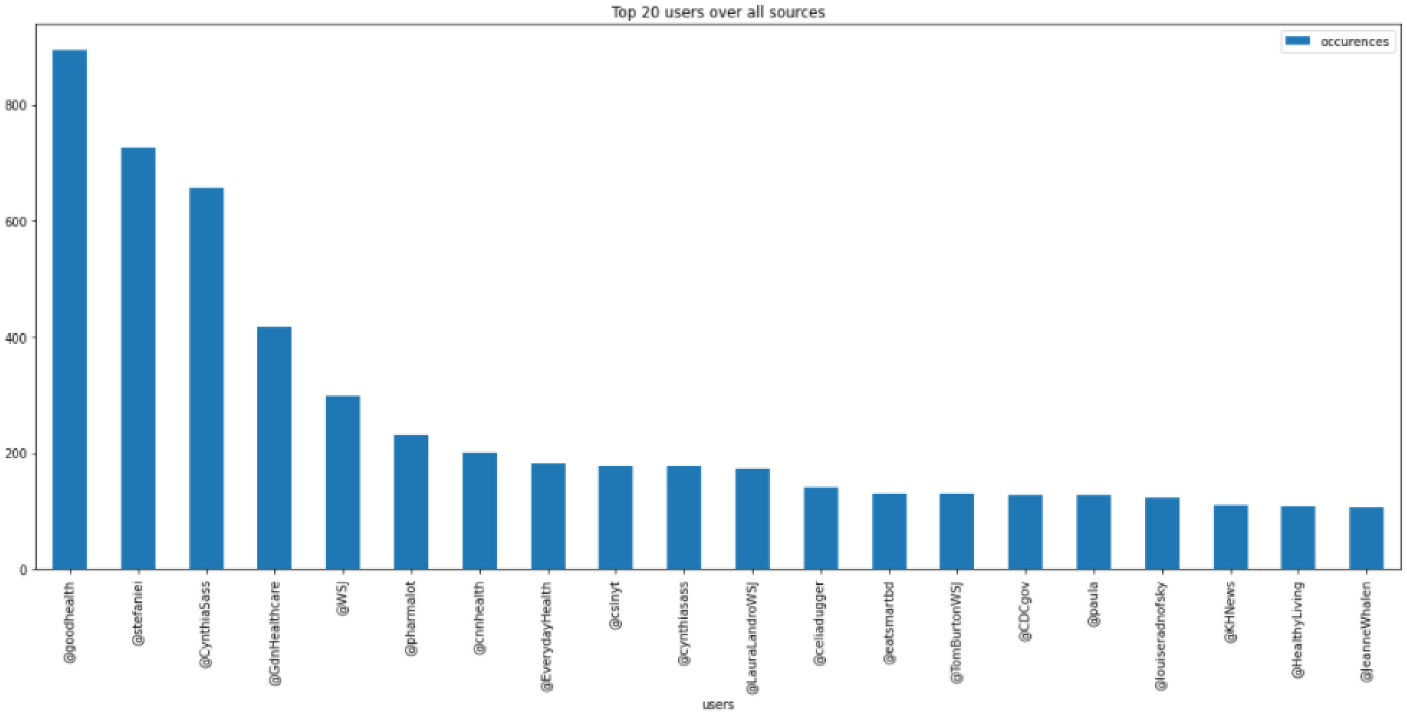


EXPLORATORY DATA ANALYSIS (@USERMENTION ANALYSIS)



	users	occurences
3138	@goodhealth	894.0
4200	@stefaniei	725.0
486	@CynthiaSass	656.0
864	@GdnHealthcare	418.0
2347	@WSJ	299.0

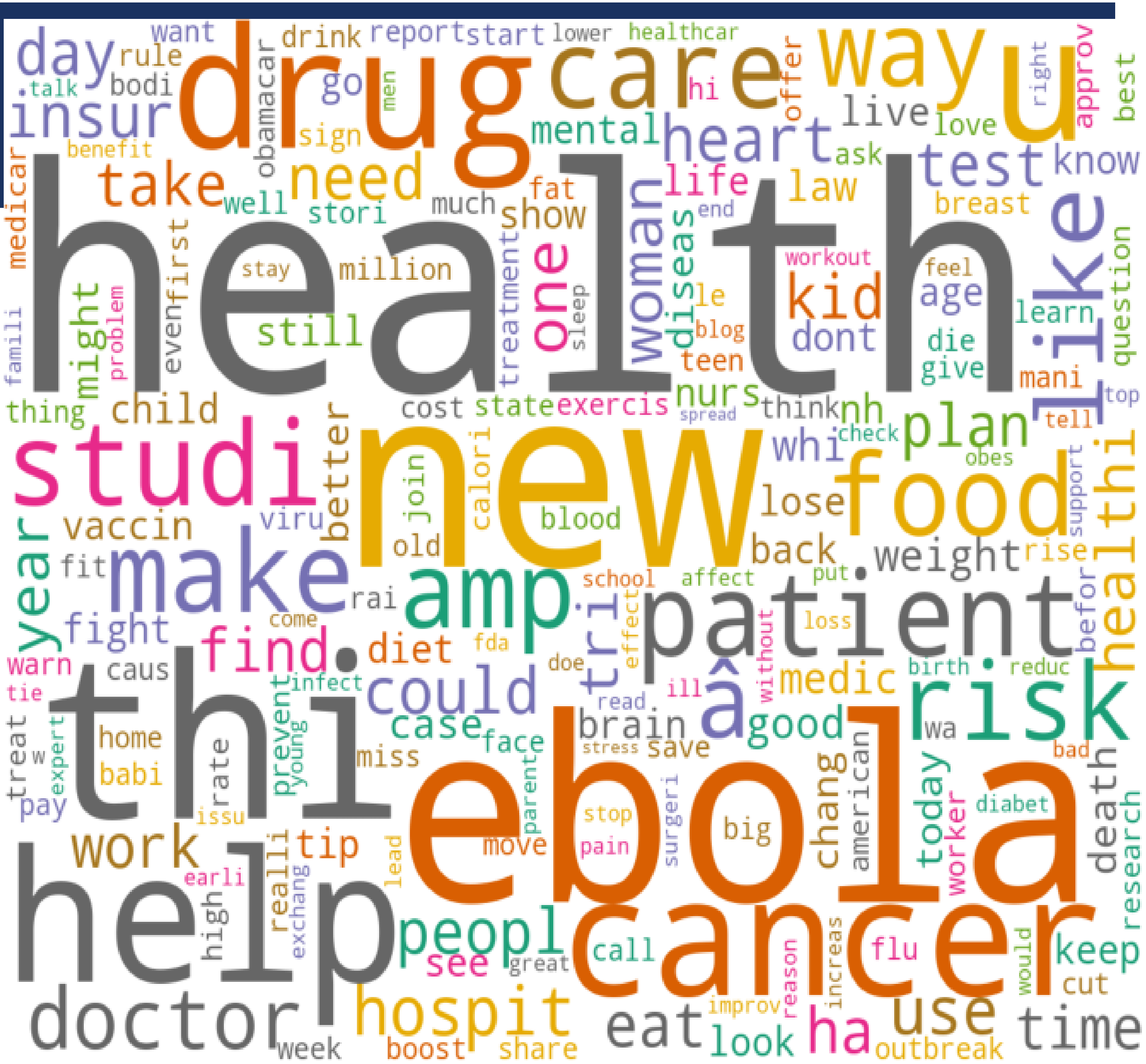
Top 20 trending Mentions across all news agencies.



EXPLORATORY ANALYSIS

Word Occurrence

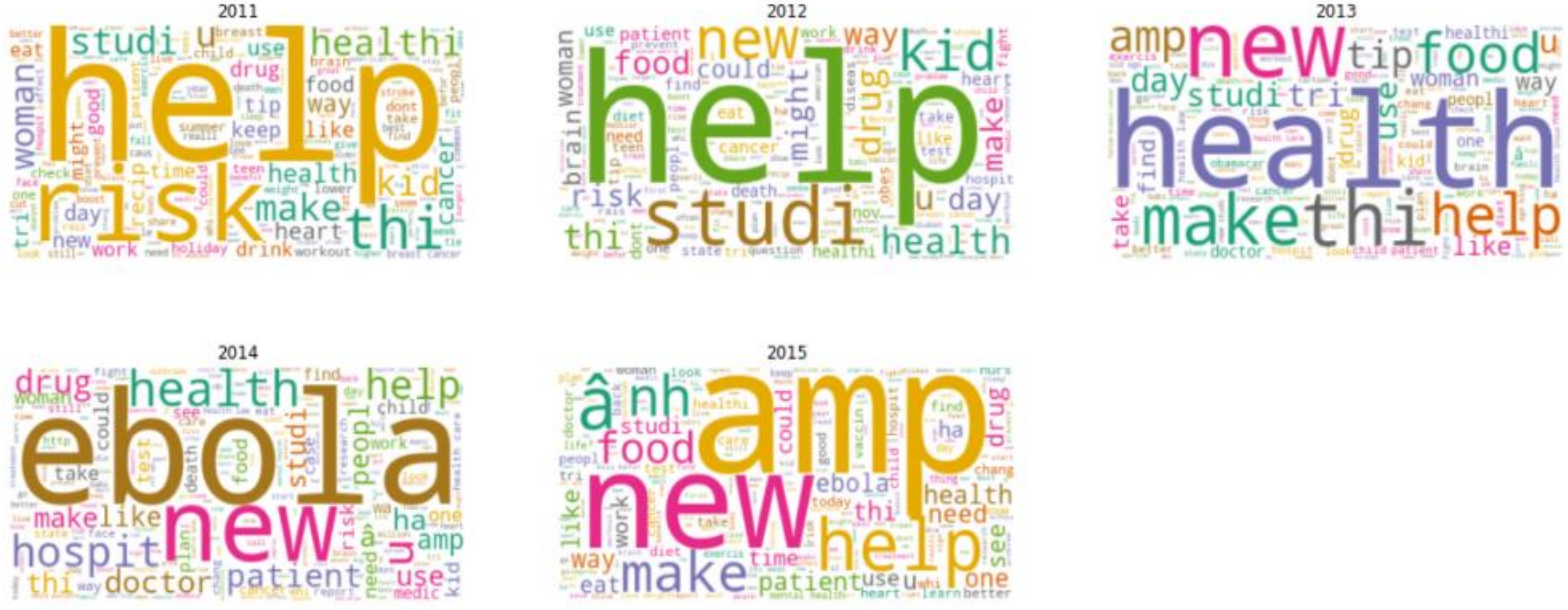
Most Tweeted Words between 2011 to 2015



EXPLORATORY ANALYSIS

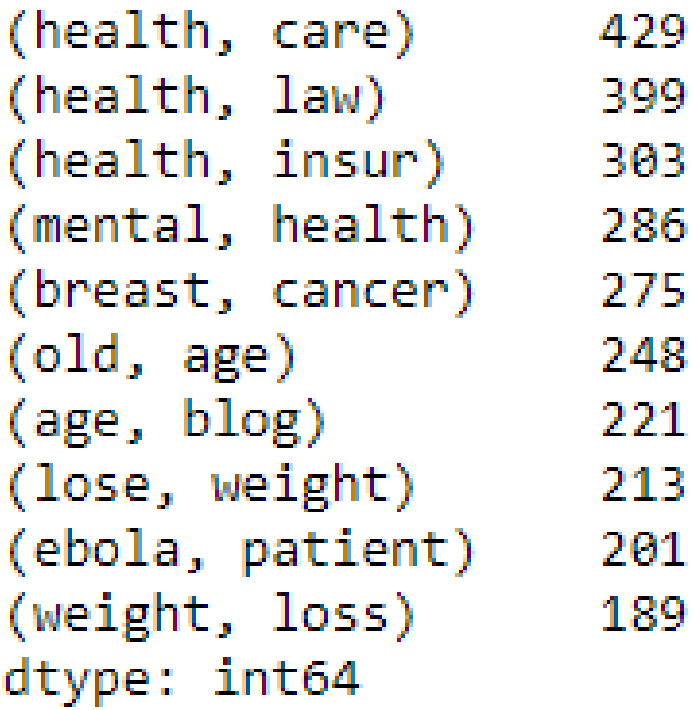
Word Occurrence

Most tweeted words by Year

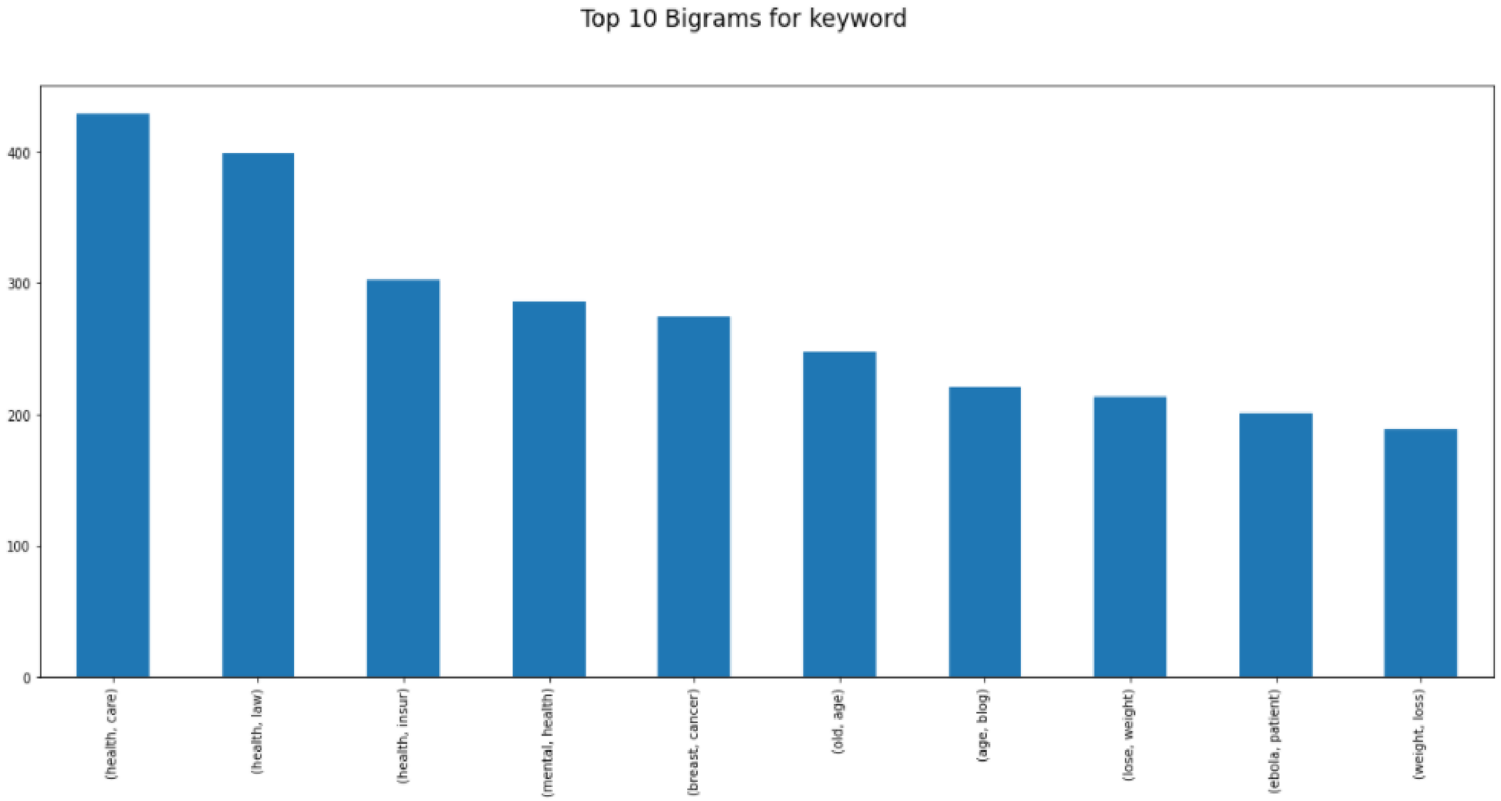

2011
2012
2013
2014
2015

EXPLORATORY ANALYSIS

Top 10 Bigrams



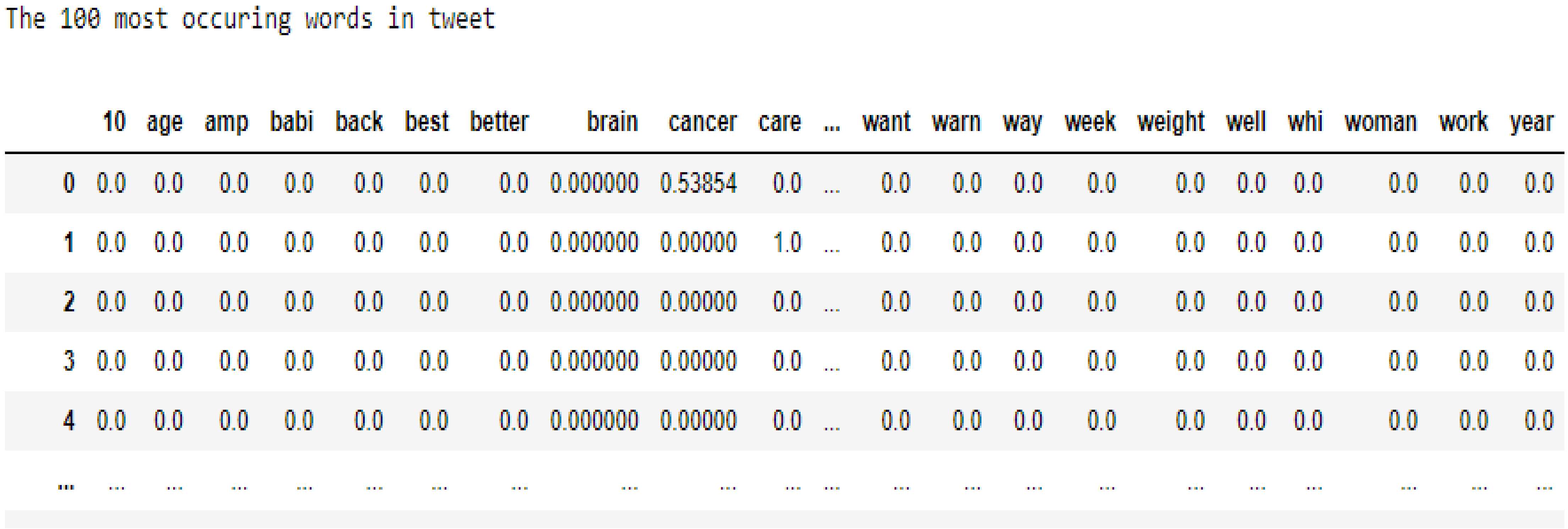
```
(health, care)        429
(health, law)         399
(health, insur)       303
(mental, health)      286
(breast, cancer)      275
(old, age)            248
(age, blog)           221
(lose, weight)        213
(ebola, patient)      201
(weight, loss)        189
dtype: int64
```



MAIN ANALYSIS (UNSUPERVISED LEARNING - DATA CLUSTERING)

Method Adopted: K-means Clustering Method

We used **Term Frequency-Inverse Document Frequency (TF-IDF)** to convert tweets to a sparse matrix of weighted frequency as shown below;



The 100 most occuring words in tweet

	10	age	amp	babi	back	best	better	brain	cancer	care	...	want	warn	way	week	weight	well	whi	woman	work	year
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.53854	0.0	...	0.0	0.0	0.0	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	0.0	0.000000	0.00000	1.0	...	0.0	0.0	0.0	0.0	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	0.0	0.000000	0.00000	0.0	...	0.0	0.0	0.0	0.0	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	0.0	0.000000	0.00000	0.0	...	0.0	0.0	0.0	0.0	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	0.0	0.000000	0.00000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

MAIN ANALYSIS (UNSUPERVISED LEARNING - DATA CLUSTERING)

Performance Metric

Elbow Method is determine optimal number of clusters K

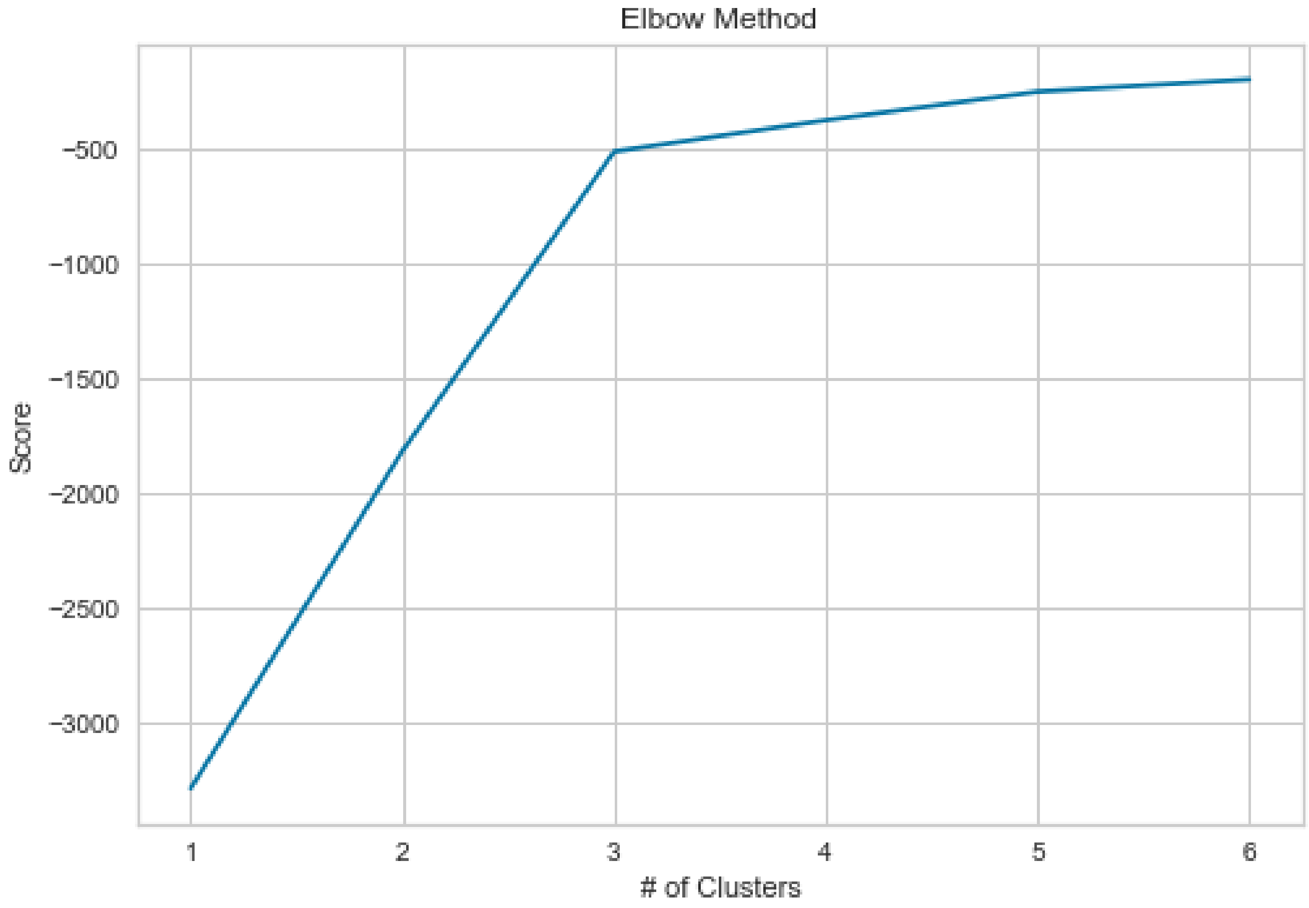
From the Image here, the optimal K clusters is 3

Also, we used the Silhouette Score to determine our optimal k clusters.

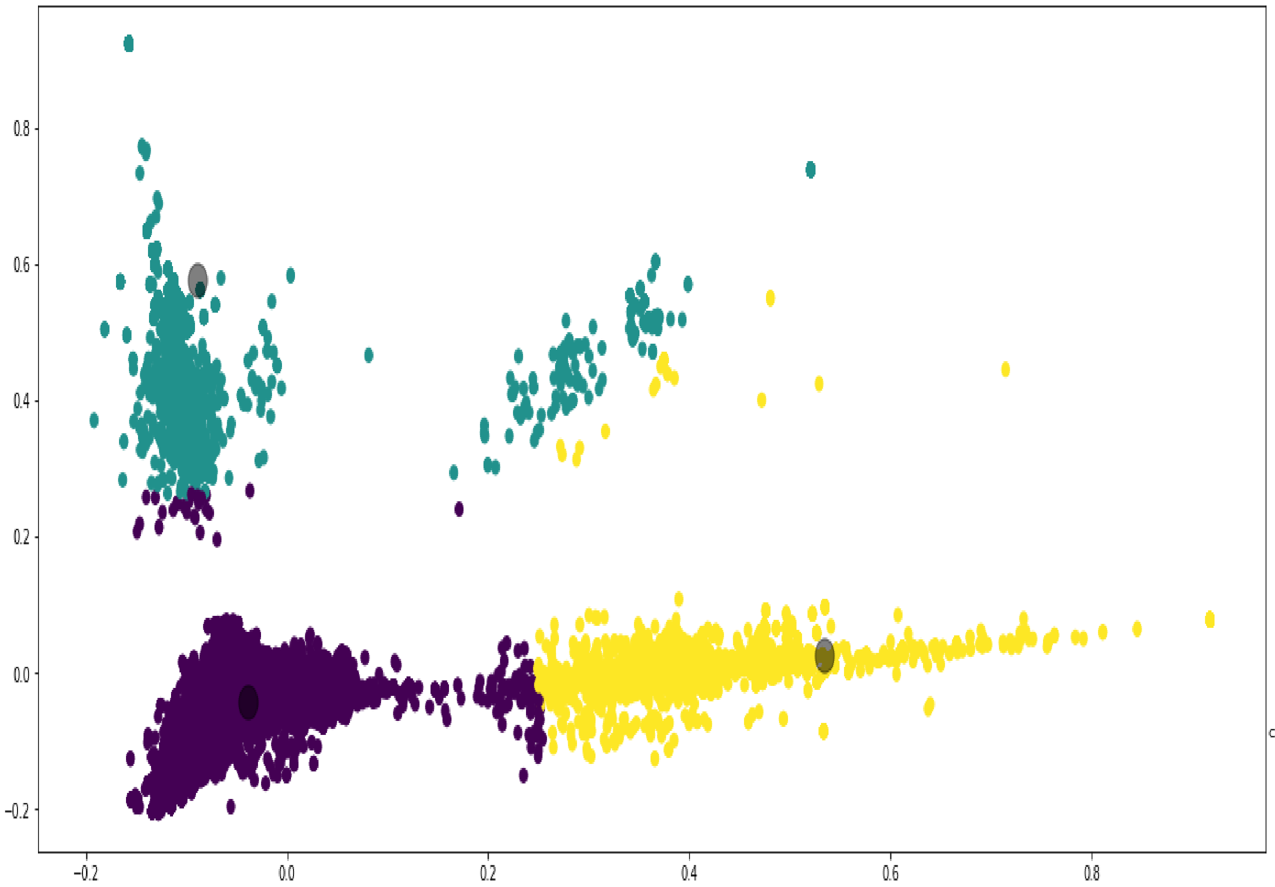
Silhouette Scores:

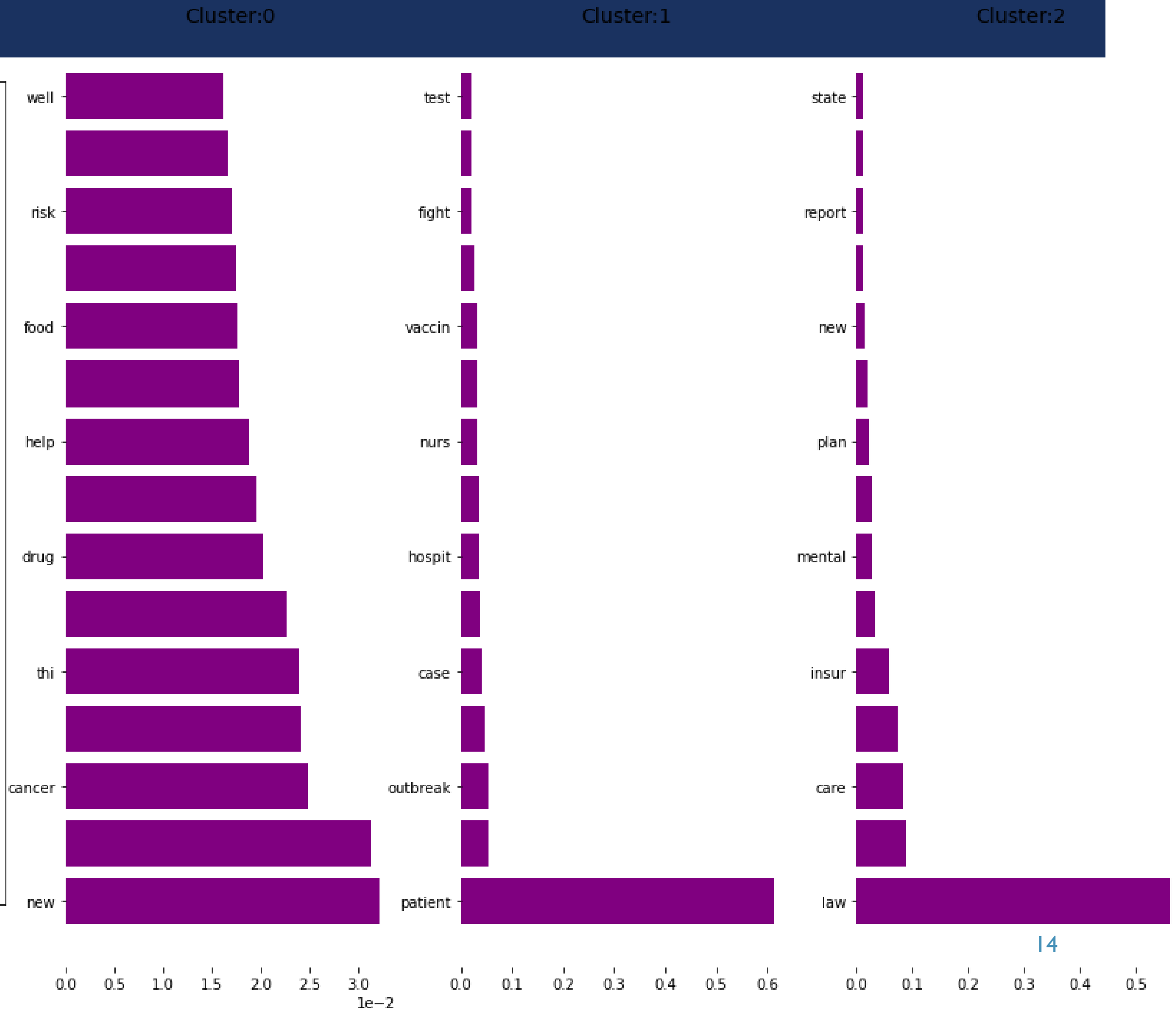
```
KMeans(max_iter=600, n_clusters=3)
Silhouette score: 0.87

KMeans(max_iter=600, n_clusters=5)
Silhouette score: 0.87
```



CLUSTERING RESULT AND SCATTER PLOT



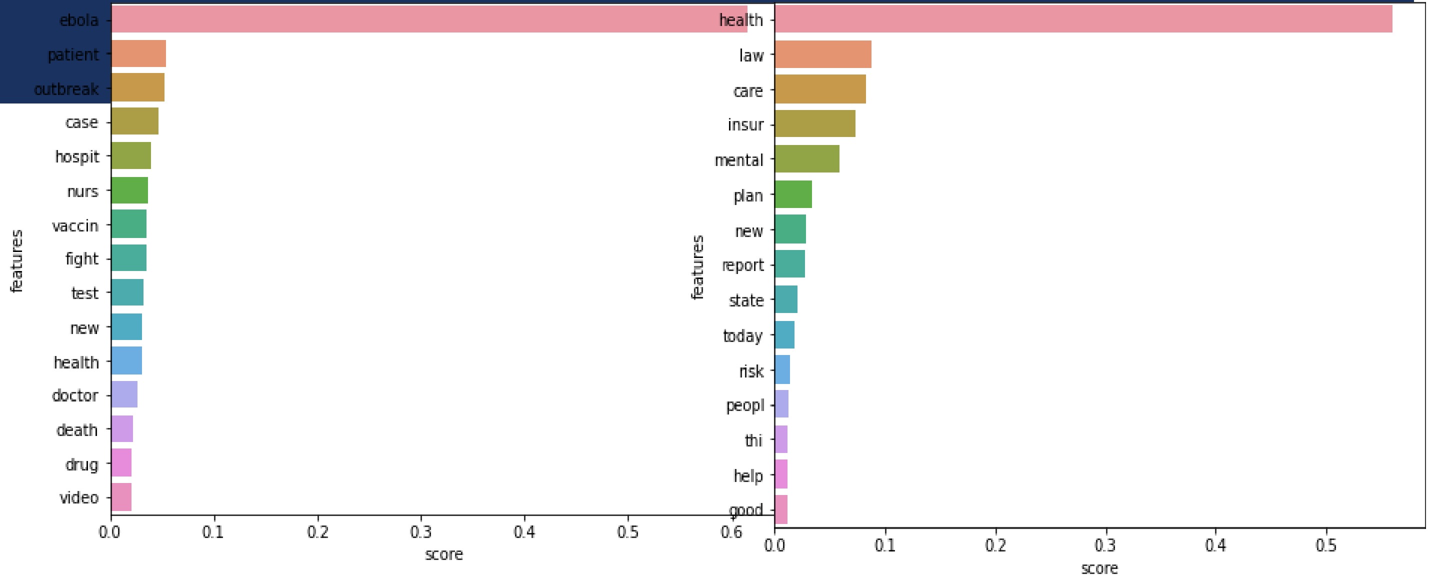


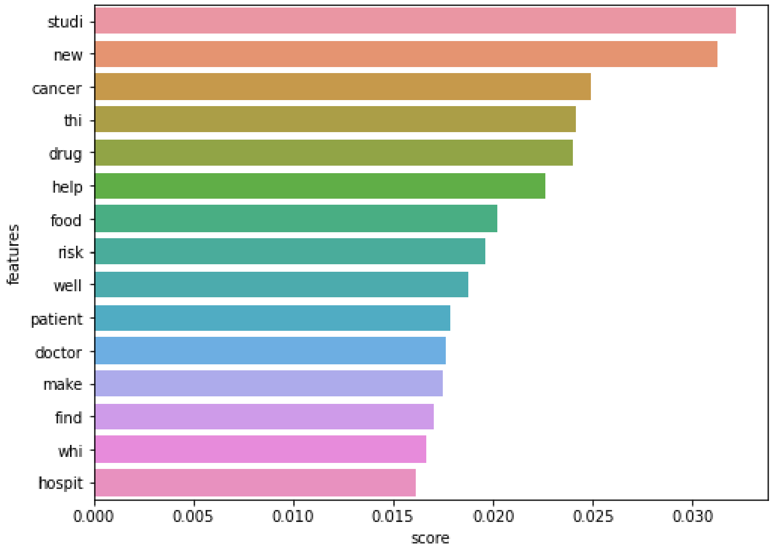
RESULT AND CONCLUSION

We can see from the three clusters;
The first cluster was centered on the well being of the patients and also studies to help people reduce the risk of having cancer

The second cluster was focused on the outbreak of the ebola disease pandemic and also studies on its vaccine.

The third cluster was concerned about health law, insurance and also the mental status of people





THANK YOU