Analysis of Tweets: Donald Trump and sexual harassment allegations

MAIN CONCEPT

Twitter historical database was used to search for suitable tweets. Search term of "Trump" and "harassment" was consequently used since term sexual harassment did not returned enough tweets for analysis. After examination of tweets connections with broader term of sexual harassment in other areas did emerge.

We used software tool NodeXL[1] which uses the Twitter API[2] connection to search Twitter database. Twitter has set so called "rate limitation" which means the search and export has limitation in quantity of tweets and users and the time frame. Twitter API allows to search only thru last 7 days of tweets history.

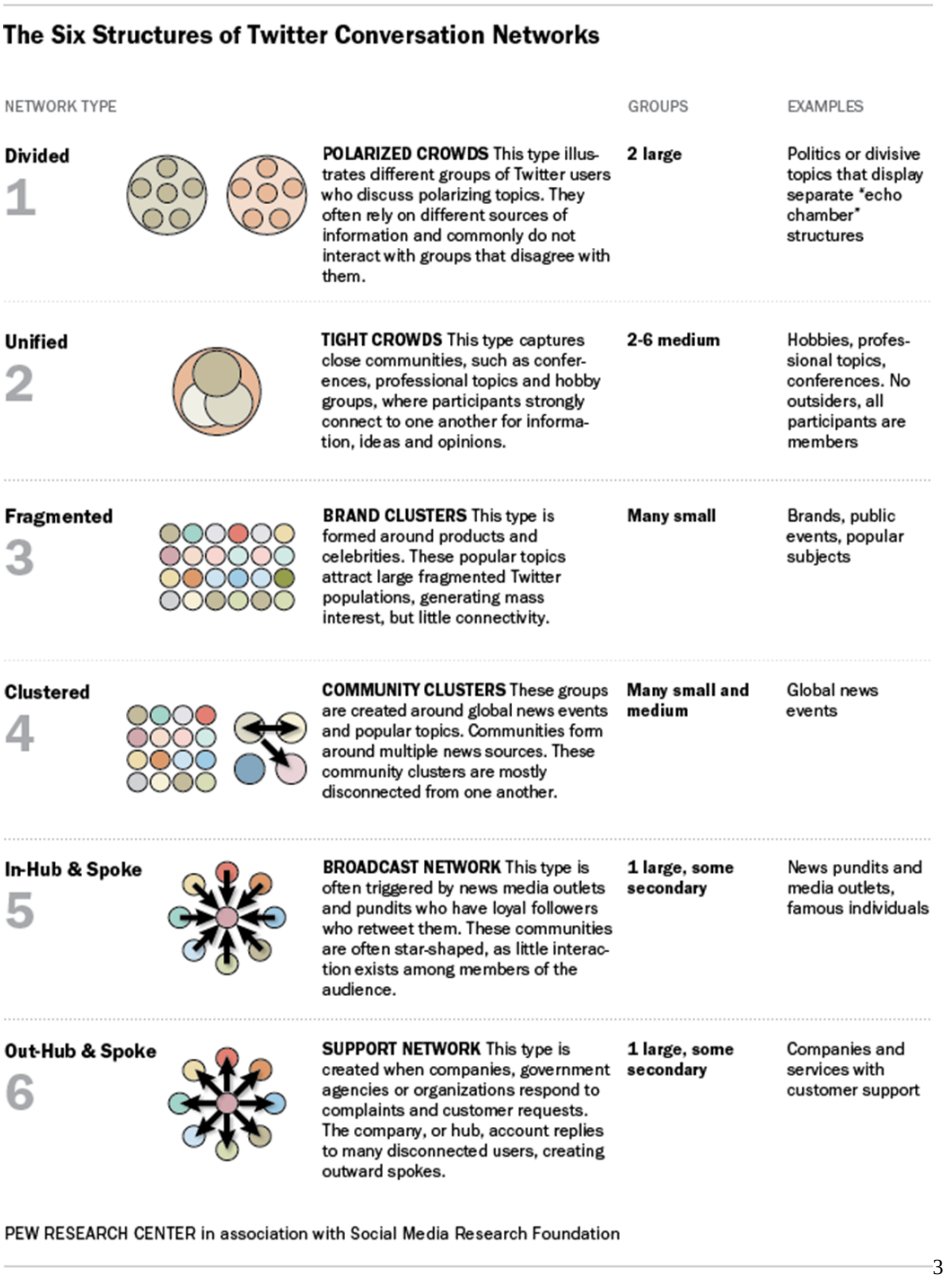
NodeXL has built-in parser to go thru search results and automatically set the proper format for network analysis and graph visualization. It does differentiate between three different type of relationship: tweet, mention and replay. Logic behind twitter is sharing, we can treat list of tweets as indicator of user activity but only the frequency of mentions and replays give us the measurement of influence of a user. It is important what kind of reach user tweets have. List of followers are important to understand how long is the reach of a user voice. This so called central or influential positions of users will be clearly identified in graphs in results section.

Network graph is formed based on lines connecting different users and representing the mentions or replays. There is also another type of network which can be drawn. Compared to graph where source of connection is actual tweets itself we can draw graphs based on users who are followers and following. In this way, we can identify user network with most important, central users and underlying clusters of similar users. Where measure of similarity is shared list of followers or following. Most commonly used in branding or marketing actions. In this way with little effort and selection of right user one can easily and quickly spread the news around. These central users are known as influencers and they have a role of being a hubs and bridges in the network.

As in any conversation and especially in social network analysis we can always identify different type of conversation. Pew Research center in association with Social Media Research

[1] http://www.smrfoundation.org/
[2] https://dev.twitter.com/rest/public

Foundation (MARC A. SMITH, 2014) has identified 6 different type of twitter conversation. This kind of differentiation is very important to understand the dynamics of network.



The Six Structures of Twitter Conversation Networks

NETWORK TYPE		GROUPS	EXAMPLES
Divided 1	**POLARIZED CROWDS** This type illustrates different groups of Twitter users who discuss polarizing topics. They often rely on different sources of information and commonly do not interact with groups that disagree with them.	**2 large**	Politics or divisive topics that display separate "echo chamber" structures
Unified 2	**TIGHT CROWDS** This type captures close communities, such as conferences, professional topics and hobby groups, where participants strongly connect to one another for information, ideas and opinions.	**2-6 medium**	Hobbies, professional topics, conferences. No outsiders, all participants are members
Fragmented 3	**BRAND CLUSTERS** This type is formed around products and celebrities. These popular topics attract large fragmented Twitter populations, generating mass interest, but little connectivity.	**Many small**	Brands, public events, popular subjects
Clustered 4	**COMMUNITY CLUSTERS** These groups are created around global news events and popular topics. Communities form around multiple news sources. These community clusters are mostly disconnected from one another.	**Many small and medium**	Global news events
In-Hub & Spoke 5	**BROADCAST NETWORK** This type is often triggered by news media outlets and pundits who have loyal followers who retweet them. These communities are often star-shaped, as little interaction exists among members of the audience.	**1 large, some secondary**	News pundits and media outlets, famous individuals
Out-Hub & Spoke 6	**SUPPORT NETWORK** This type is created when companies, government agencies or organizations respond to complaints and customer requests. The company, or hub, account replies to many disconnected users, creating outward spokes.	**1 large, some secondary**	Companies and services with customer support

PEW RESEARCH CENTER in association with Social Media Research Foundation

[3]

[3] http://www.pewresearch.org/fact-tank/2014/02/20/the-six-types-of-twitter-conversations/

NODEXL PROCEDURE

NodeXL is specialized social network analysis tool and it is an Excel add-in service. It has the capability to create, analyze and visualize different type of networks.

We will use build-in Twitter API service which connect us directly to a Twitter database. In search box, we entered our terms: trump and harassment.

Output does create two type of tables. Vertices and edges, first being the list of all users found in tweets and the edges being the connections between this user. Connections are actual content of the tweets.

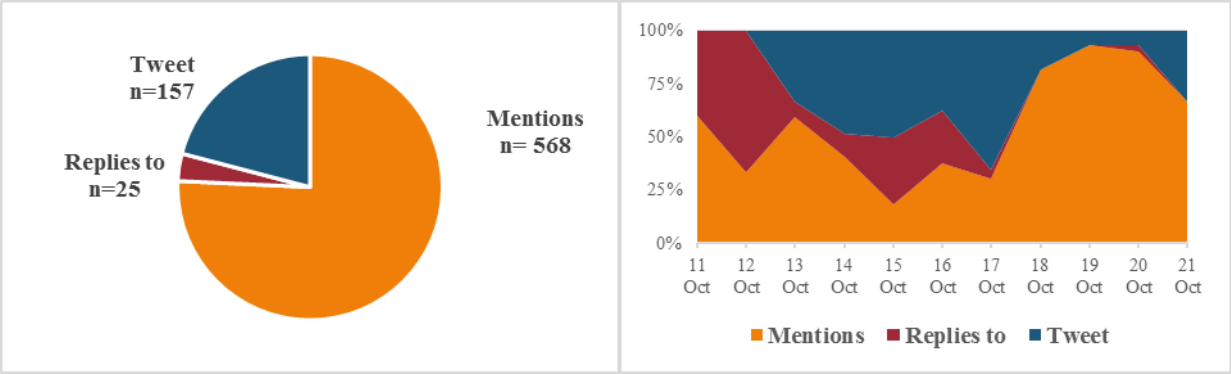
These two tables are our data sources to create a network graph. Vertices sometimes called nodes are users (dots) and edges are lines (mentions and replays) connecting different users. After the graph drawing the actual social network analysis begin. With different types of graph layout algorithms, we can search the best visual representation of our network. We can also apply different type of statistics to identify clusters, graph density, ranking, centrality, connectivity.

We can apply different visualization techniques based on metrics to create graph. We can enlarge dots based on number of connections (in or out) each user has or number of followers. We can color nodes to identify clusters or to add different user attributes. We can enlarge edges (lines) to emphases the importance of connections for this we can use different type of weights which represents the power of connection between the users.

RESULTS

Our sample has list of tweets from 11.10.20016 until 21.10.2016. There are 757 users (vertices) and 752 connections (edges) and 750 unique connections. This graph is directed[4] graph and has 157 self-loops or as we can identify them as tweets with no replay or mentions.

Basic structure and statistic of our tweet sample:



Time is also important dynamic factor for network growth. We can see how time affected increase in number of mentions during, between and short after last presidential debate on 19.10. If mentions are passive components, replies and tweets are active. It is obviously that active components grow after passive component take care for distribution/sharing. Previous debate was on 9.10 and after few days, shares of active components are bigger. Later we will show dynamic component in network layout.

We counted 157 tweets, 568 mentions and 25 replies. Graph has 175 closed subgraphs; these are so called connected components and in our case these are users connected with each other with two type of connections (mentions and replies). It also has 121 single users which are not connected with anyone. Maximum size of a subgraph is 357 users and maximum number of connection in a subgraph is 361. The maximum number of connections to get from one user to another is 8 with the average number of 3,2 connection. Our sample graph is a very sparse graph. Maximum density of a graph is 1 and minimum is 0, our is 0,00104 which lead us to the conclusion that sample of users sharing same search term are not closely connected. We will show this in graph visualization.

[4] A graph whose edges are ordered pairs of vertices. Each edge can be followed from one vertex to another vertex.

Graph Type	Directed
Vertices	755
Unique Edges	750
Edges with Duplicates	0
Total Edges	750
Self-Loops	157
Connected Components	175
Single-Vertex Connected Components	121
Maximum Vertices in a Connected Component	357
Maximum Edges in a Connected Component	361
Maximum Geodesic Distance (Diameter)	8
Average Geodesic Distance	3.217213
Graph Density	0.00104168

In next table, we are presenting most used graph metrics which will enable us to recognize structure of our network.

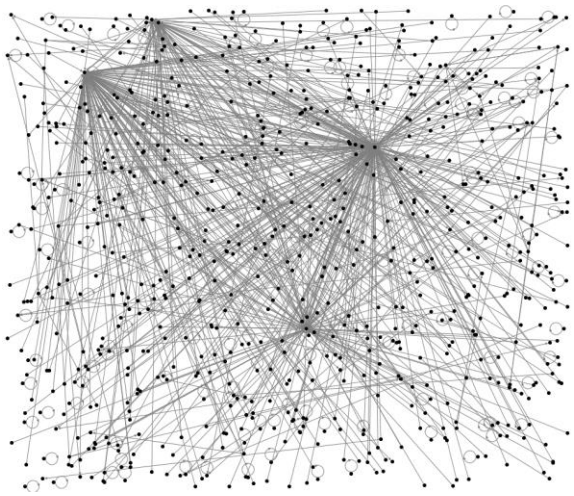
	Minimum	Maximum	Average	Median
In degree	0	185	0.993	0
Out degree	0	5	0.993	1
Betweenness Centrality	0	106614	401.446	0
Closeness Centrality	0	1	0.1	0.001
Eigenvector Centrality	0	0.07	0.001	0
PageRank	0.435	84.620	1	0.544
Clustering Coefficient	0	0.667	0.01	0

With in/out degree we are recognizing the most influential or popular user. Betweenness centrality tells us which users are mostly likely connected with other users. Closeness gives us the shortest path or speed to reach other users from a starting user. Eigenvector centrality help us recognize who is connected to the most connected users. Page rank gives more importance to connection, connected to the most important users. Clustering coefficient is the number of closed triplets and it help us recognize the clusters. Above are aggregated metrics but in detail each user has own metric calculated.

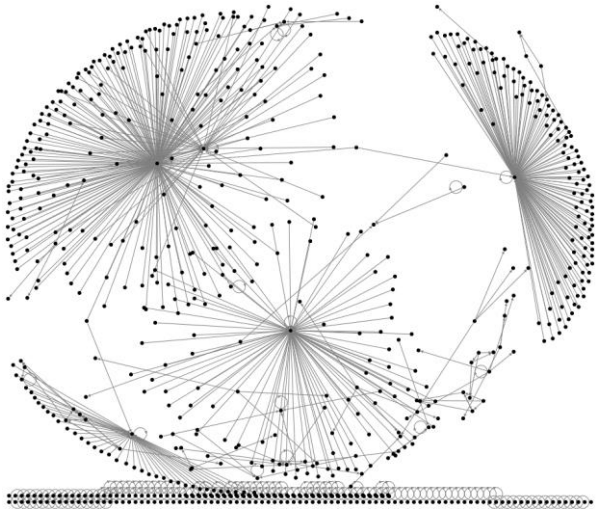
VISUALIZATION

In this basic graph visualization[5], it is hard to recognized network structure but few central users could be identified. In another instance, we did adjust layout and there are clusters visible and several smaller groups and users as loops.

Draft layout



Boxed layout

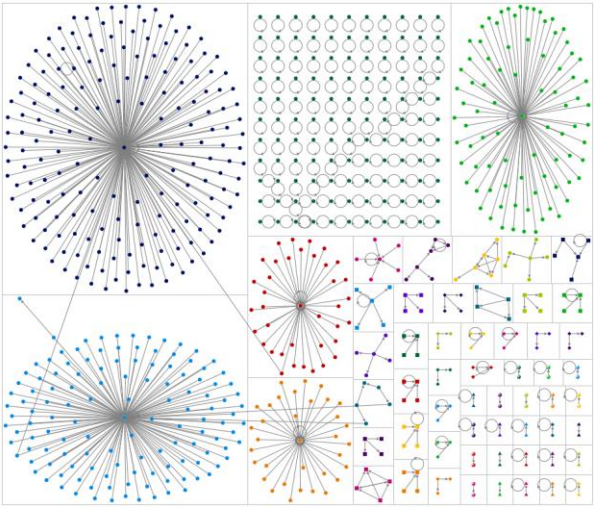


For better understanding we created a clustered layout so the structure of our network is easier to read. Below is the graph with applied clusters and different layout where we can identify large number of isolated subgraphs and three large centered subgraphs. Loops or as we saw single tweets have no visual value for us. We will filter them out and focus on groups with more connections.

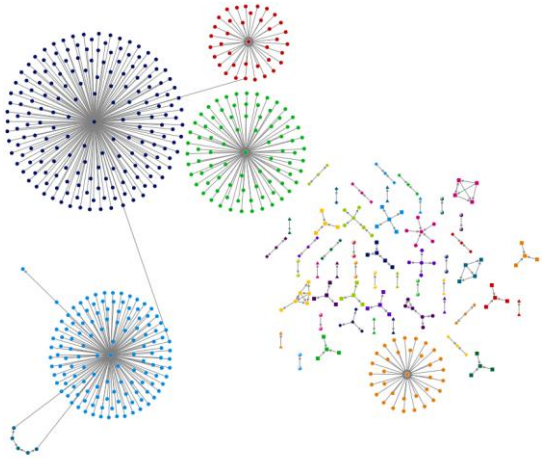
[5] We used Fruchterman-Reingo layout algorithm.

After applying filter on relationship type and use only mentions and replays we are getting more clear visual representation of our network based on selected search term.

Clustered layout



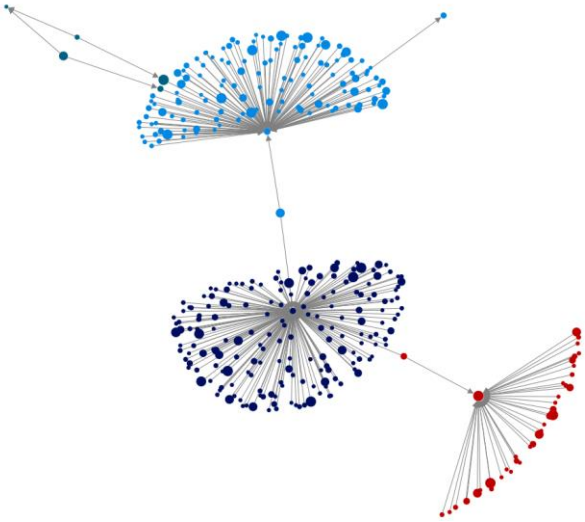
Clustered layout without the loops



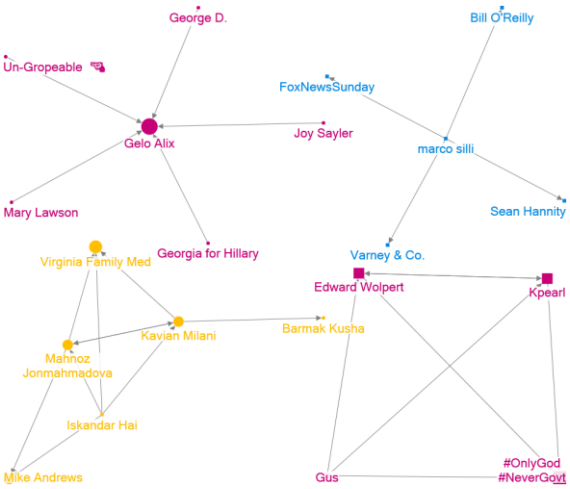
We can recognize our network as clustered network, with many divided groups but within this clusters most of the large ones are broadcast network type. Between big centered graphs we can recognize few users with the role of bridge between different centered graphs. This is important

Connected subgraph[6] with central users

(node size = historical number of tweets)



Isolated subgraphs with different structure and size of nodes as in-degree value



for sharing information to a wider spread of users as those users are vital for network growth.

[6] We used Harel-Koren Fast Multiscale layout alghoritm to expose bridge users

We can apply different metric or attributes to nodes or edges this kind of graph feature gives us another perspective of graph visualization and better understanding of underlying structure. In case of tweet network, it is important the actual reach certain user has in perspective of the number of followers and followed. Centrality measures are here vital to understand the “power” of certain user. Like in any analysis it is important to view all available metric to avoid ambiguity.

Most of our cluster are broadcast networks which means they are star shaped and in central position we often have famous individuals or organizations with large number of followers which are not closely connected.

GROWTH OF NETWORK THRU TIME

As we said dynamic component of network growth is highly dependent from the time of the news/topic/event is entered in a tweeter space. Network is growing in time but the momentum of growth is highly dependent from the influential users.

11-12.10.2016

12-13.10.2016

13-14.10.2016

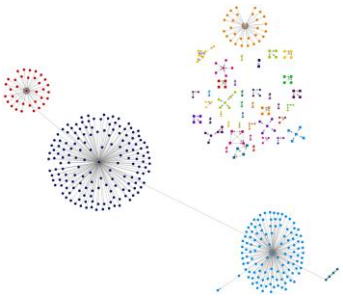
14-15.10.2016

15-16.10.2016

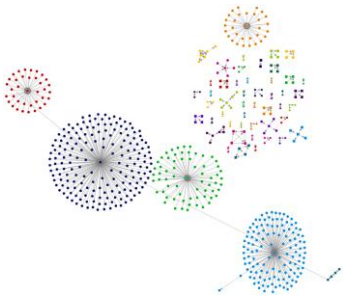
16-17.10.2016

17-18.10.2016

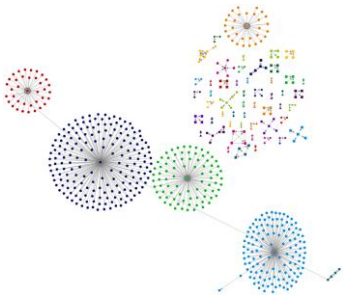
18-19.10.2016



19-20.10.2016



20-21.10.2016

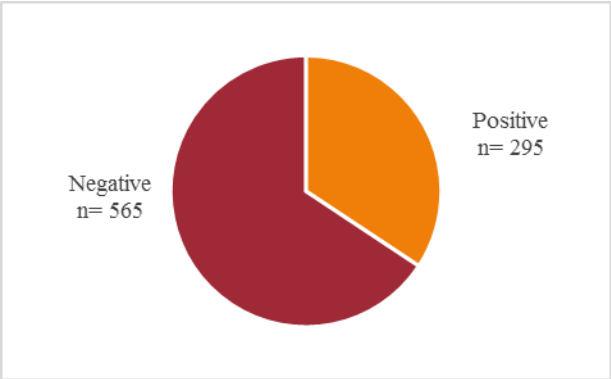


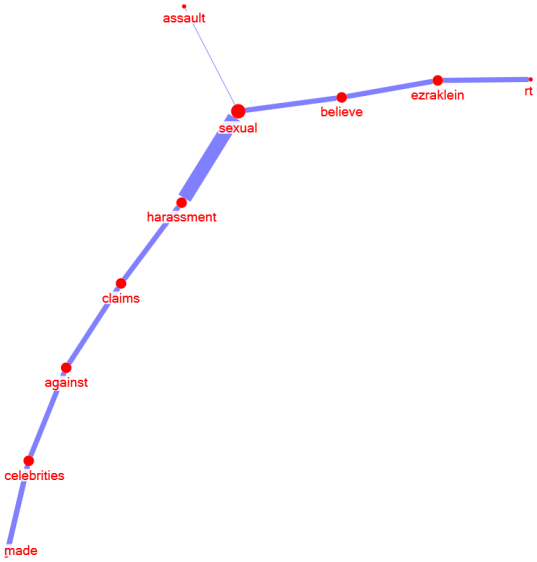
CONTENT ANALYSIS

Social network analysis can give us an insight of relationship structure and dynamics. There is also a content analysis which can be in case of our Tweeter case of vital importance to understand in which direction debate is developing. Simple content analysis is known as sentiment analysis:

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker. A common use case for this technology is to discover how people feel about a particular topic. (https://www.lexalytics.com/technology/sentiment)

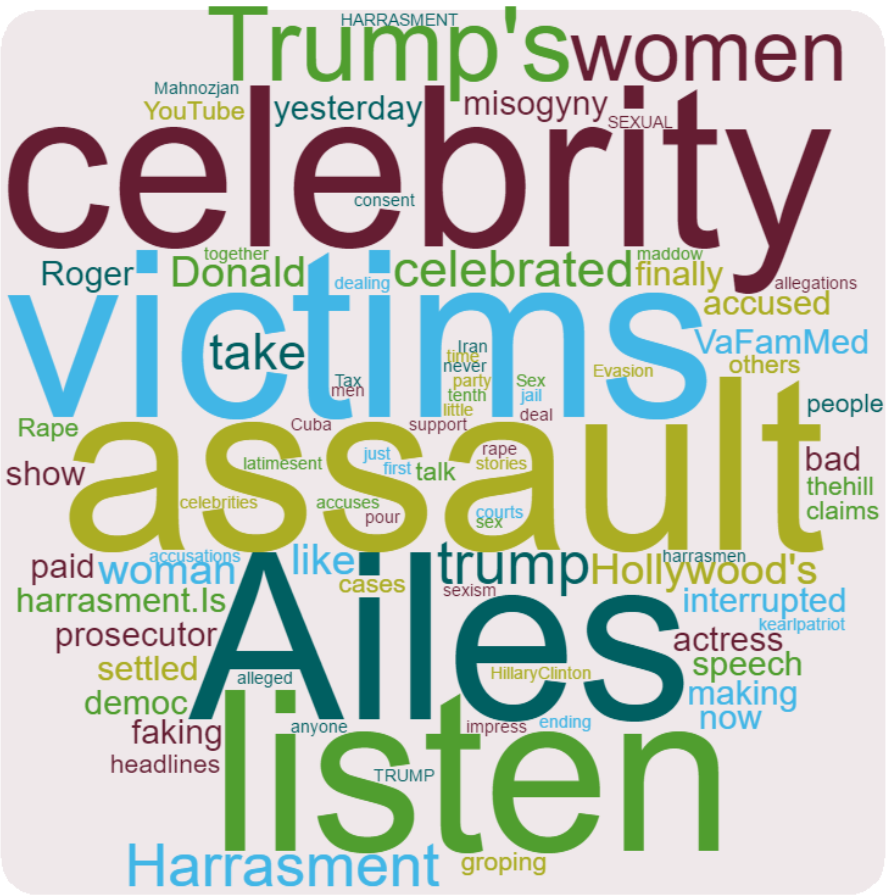
In our case sentiment analysis showed us that our search term returned tweets with more negative sentiment. The outcome could be different if all the words would be categorized. Software default did manage to classify 860 words while total word count is 9161. Even if not all world could be categorized sentiment analysis could be different.





Another perspective of content analysis is word pairs so you can see which word are mostly used together in a tweet. This is showing only top 10 pairs and unfortunately we can't see if our first search term "trump" would also connect with any words mention in graph above.

Perspective which can be identify based on content analysis is also how debate could be spread in other areas of public life. If Donald Trump allegations of sexual harassment were the starting point of the debate, there is even more interesting development in direction of overall problem of harassment. A simple world cloud[7] show as the width of the problem.



[7] http://www.wordclouds.com/

TWEET CONTENT ANALYSIS

In this final stage, we will present Top network items.

Top URLs in Tweet in Entire Graph	Entire Graph Count
http://www.vox.com/identities/2016/10/18/13306300/trump-cosby-ailes-sexual-harrasment?utm_campaign=vox&utm_content=chorus&utm_medium=social&utm_source=twitter	47
http://www.vox.com/identities/2016/10/18/13306300/trump-cosby-ailes-sexual-harrasment/in/13029661?utm_campaign=vox.social&utm_medium=social&utm_content=voxdotcom&utm_source=twitter	30
http://www.latimes.com/entertainment/tv/la-et-st-sexual-harrasment-hollywood-trump-bush-20161012-snap-story.html	15
http://www.vox.com/identities/2016/10/18/13306300/trump-cosby-ailes-sexual-harrasment?utm_campaign=vox&utm_content=entry&utm_medium=social&utm_source=twitter	11
http://trib.al/P6AHLQN	10
http://www.vox.com/identities/2016/10/18/13306300/trump-cosby-ailes-sexual-harrasment	10
http://www.latimes.com/entertainment/tv/showtracker/la-et-st-sexual-harrasment-hollywood-trump-bush-20161012-snap-story.html?utm_source=dlvr.it&utm_medium=twitter	9
http://route.overnewser.com/politicsnewz/?url=http://www.latimes.com/entertainment/tv/la-et-st-sexual-harrasment-hollywood-trump-bush-20161012-snap-story.html&utm_source=twitter-tools&utm_medium=PoliticsNewz&utm_campaign=article	5
http://www.latimes.com/entertainment/la-et-st-sexual-harrasment-hollywood-trump-bush-20161012-snap-story.html?utm_source=dlvr.it&utm_medium=twitter	5
https://www.theguardian.com/commentisfree/2016/oct/14/michelle-obama-us-election-donald-trump-politics-speech-women?CMP=fb_gu	5

Top Domains in Tweet in Entire Graph	**Entire Graph Count**
vox.com	134
latimes.com	53
trib.al	5
youtube.com	3
overnewser.com	2
theguardian.com	2
linkedin.com	1
huffingtonpost.com	1
twitter.com	1
nippysky.com	1

Top Hashtags in Tweet in Entire Graph	Entire Graph Count
trump	2
strongwomen	2
harrasment	2
misogyny	2
sexism	2
gold	1
forex	1
debate	1
women	1
bloggersblast	1

Top Words in Tweet in Entire Graph	Entire Graph Count
Words in Sentiment List#1: Positive	295
Words in Sentiment List#2: Negative	565
Words in Sentiment List#3: (Add your own word list)	0
Non-categorized Words	8301
Total Words	9161
sexual	533
rt	503
celebrities	368
harassment	277

Top Word Pairs in Tweet in Entire Graph	Entire Graph Count
sexual,harassment	277
rt,ezraklein	184
ezraklein,believe	184
believe,sexual	184
harassment,claims	184
claims,against	184
against,celebrities	184
celebrities,made	184
made,celebrities	184
sexual,assault	137

Top Replied-To in Entire Graph	Entire Graph Count
skynews	1
rosie	1
ebako_e	1
beeplus22	1
benjysarlin	1

tmsp2003	1
colony14	1
franks_3111	1
hillaryclinton	1
drudge_report	1
Top Mentioned in Entire Graph	**Entire Graph Count**
ezraklein	184
voxdotcom	131
hectorguilleng	83
latimes	37
lorraineali	33
geloalix	5
youtube	3
ihollaback	3
vafammed	3
mahnozjan	3

Top Tweeters in Entire Graph	**Entire Graph Count**
wrikent3500	707194
marekingu	577130
missnisha6849	452825
rjbailey	439280
rwneilljr	396227
sumersloan	395115
lawsonbulk	388170
foxnews	274927
paxnostrum	248413
thehill	240452

CONCLUSIONS

US elections are always hot world topic with all the allegations and bad publicity. We expose very delicate matter of sexual harassment which is a worldwide problem where victims are not heard and public was and still is very disturbed by the behavior of presidential candidate Donald Trump.

It is not our intention to recognize the truth behind allegations but the fact is that users not only condemned the behavior of Donald Trump but also expose connected problems with harassment. It is a wider problem present in every pore of our society and the victims are not heard.

In this example, we showed that simple content analysis of a social network can give use the meaning and every attributes or metric can be used to enhance the graphical visualization.