CmpE 58H - Social Semantic Web
Fall 2015

Final Project Report
Network Analysis on Twitter Data

Nihal Yağmur Aydın
Hakime Öztürk

Boğaziçi University,
2016

1. INTRODUCTION

Twitter data is like many other social media data, noisy, informal, and includes many slang words, so it is difficult to process tweets semantically [1].

Even with its compressed way of communication, Twitter data promises valuable information to be extracted such as profiling user types. The word-user networks provide social insights about the featured context as well as revealing different personas involved. Because of that, we used Twitter data, considering word-pairs centered around the slang words to reveal information regarding users current interests. Our topic focuses on “Trump” keyword which denotes to the politician from Republicans Party of USA, Donald Trump.

In this project, we first define a slang ontology for Twitter and then, extract a number of semantic relations between the words and words, and words and users. We then visualize the mined relations as networks of different relations to investigate possible interactions and user profiles that an analysis can reveal.

2. MODEL

Social media, connecting millions of people each second, provides invaluable source for knowledge elicitation. However microblogs, such as Twitter, force it users to conceive their ideas with limited number of words and sometimes be as fast as possible to publish what they think. Because of this strict environment, emergence of a social media language is inevitable. Among the major components of the social media language, use of abbreviations (short-forms and acronyms) has the significant percentage along with the typing errors/misspells, emoticons and wordplays [2].

In the model we propose, first tweets about a certain topic are collected. Then, the tweet streams are parsed so that the resulting data is in the following format: username, tweet text, user location, and the language (i.e., username: “lee_lizard”, tweet: “@MrDanRigby nope not a thing I do it all on Xmas eve!!” , location: “Liverpool”, language: “en”).

After that, the processed data is filtered so that only the tweets containing slang words and written in english (e.g., language: “en”) are remained in the data set. We have used an online slang word dictionary [7], which comprises 237 different slang words, in order to identify slang words in the tweet text (e.g., 20: location, ne1: anyone, tnx: thanks etc.).

The data processing and creating the semantic relationships are completed using Java programming language. The source code and the input files used in this project is available in our GitHub page (https://github.com/sb-b/SSW-TweetProject) as well as the output files that are mentioned thoughtout the article.

2.1 Ontology

In Twitter environment, users use both slang and normal words while tweeting. We are interested in the relations between users and the words they prefer to use and the relations between slang words and the words they co-occur with in a tweet text.

There are mainly three actors in our system, which are user, slang word and real word. These actors and their relations cannot be fully represented by a simple graph with a uniform node and a uniform edge. We need several types of nodes and edges to fully represent such a system. For this reason, we have modeled our system using an ontology. Figure 1 illustrates the main structure of our slang ontology.

The slang ontology defines the possible relationships between a user and a word/slang as well as between a slang and a word in Twitter environment.

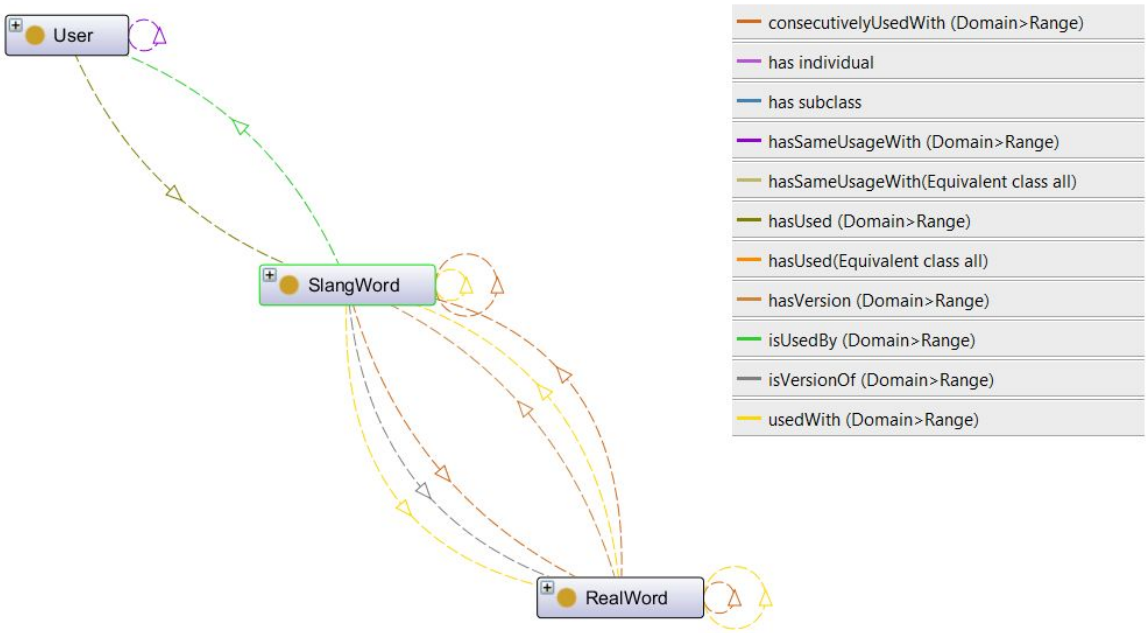

Figure 1: Basic Structure of Slang Ontology

We also tried to define our data with RDF, by using Jena library of Java, allowing creation of RDF. Below Figure 2 shows an example of such creation:

```
<rdf:RDF
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:vcard="http://www.w3.org/2001/vcard-rdf/3.0#">
  <rdf:Description rdf:about="http://localhost/">
        <vcard:N rdf:parseType="Resource">
        <vcard:Other>ok</vcard:Other>
        <vcard:NAME>had</vcard:NAME>
        </vcard:N>
        <vcard:FN>usedWith</vcard:FN>

  </rdf:Description>
</rdf:RDF>
```

Figure 2: Sample RDF

Jena library allows also querying on ontologies from Java programs and we also tested it with our ontology.

2.2 Social Network Analysis

For the network analysis, we looked at the interactions regarding user-slang, slang-co-occurrence, slang consecutive-co-occurrence and user-location. That bring the issue of semantically representing data in an appropriate form. Because of that, we modeled our data with the relations defined in the ontology.
For co-occurrence network models, we experimented the data with/without stop words included to observe how it affects the network topology. Stop words are the frequent words such as the, like, to used in the text which does not improve the specificity of the context but required for grammatical purposes.

2.2.1 Network Types

2.2.1.1 Slang-Word Co-Occurrence Network
This network represents the relationship between the actual words and the slang words that are used in the same tweet. The nodes of the network are either slang words or actual words that are connected with the 'co-occurrence' relationship. In the ontology, this relation is defined as *usedWith*.

2.2.1.2 Slang-Word Consecutive Co-Occurrence Network
In this network model, the slang words are connected with the words that appear before and after them in a tweet stream. In the ontology, *consecutivelyUsedWith* relation is used to define this interaction.

2.2.1.3 User-Slang Network
When a user uses a slang word in his/her tweet, the slang words and the user are connected in the network. There are two different node types in this network, namely, user and slang which are connected through *hasUsed/isUsedBy* relationships as they are defined in the ontology.

2.2.1.4 Location-Slang Network

This network contains relationships between slang words and the location of the users that use them.

2.2.2 Network Analysis Elements

In this section the network analysis elements are introduced.One of them is the number of components which are connected. If the components are less connected, it shows that there is a strong connectivity in general.

2.2.2.1 Clustering Coefficient

"In undirected networks, the clustering coefficient C_n of a node n is defined as $C_n = 2e_n/(k_n(k_n-1))$, where k_n is the number of neighbors of n and e_n is the number of connected pairs between all neighbors of n" *[3].*

In a more simpler form, clustering coefficient is N/M, where N represents number of edges between neighbors of n, with M being the possible maximum number of edges between neighbors of n. So, higher ratios for clustering coefficient shows the higher number of edges between neighbors of n.

2.2.2.2 Closeness Centrality

"The closeness centrality $C_c(n)$ of a node n is defined as the reciprocal of the average shortest path length and is computed as follows:

$C_c(n) = 1 / avg(L(n,m))$,

where $L(n,m)$ is the length of the shortest path between two nodes n and m." [3]

Isolated nodes in the network has the closeness centrality 0. Closeness centrality is a measure of how fast information can go from one node through the network of nodes.

2.2.2.3 Degree Centrality

"In undirected networks, the node degree of a node n is the number of edges linked to n" *[3].*

So, higher values of degree centrality shows more edges rooting from a node.

2.2.2.4 Betweenness Centrality

"The betweenness centrality $C_b(n)$ of a node n is computed as follows:

$C_b(n) = \sum_{s \neq n \neq t} (\sigma_{st}(n) / \sigma_{st})$,

where s and t are nodes in the network different from n, σ_{st} denotes the number of shortest paths from s to t, and $\sigma_{st}(n)$ is the number of shortest paths from s to t that n lies on." [3]

Betweenness centrality is about the control that a node has over other nodes. Higher values of betweenness centrality shows that dense subnetworks are connected to each other by the node.

2.2.2.5 Eccentricity

“The maximum non-infinite length of a shortest path between n and another node in the network. If n is an isolated node, the value of this attribute is zero.” [3]

So, eccentricity shows that how far a node can be connected to any other node in the network.

2.2.2.6 Average Shortest Path

“The length of the shortest path between two nodes *n* and *m* is *L*(*n*,*m*). The network diameter is the maximum length of shortest paths between two nodes.” [3]

The average shortest path length gives the value of the expected distance between two connected nodes.

2.2.2.7 Markov Clustering Algorithm (MCL)

MCL is a network clustering algorithm with a design based on mathematical bootstrapping procedure. “The process deterministically computes (the probabilities of) random walks through the sequence similarity graph, and uses two operators transforming one set of probabilities into another. It does so using the language of stochastic matrices (also called Markov matrices) which capture the mathematical concept of random walks on a graph.” [4]

2.3 Case Study: Twitter “Trump” Data

In order to demonstrate our model, we provide a case study in which the keyword “Trump” is used to collect tweet texts. Trump keyword associated with the American politician Donald Trump who is one of the strong and interesting candidates of the upcoming 2016 US president elections. Figure 3 shows sample tweets containing the word “Trump” with slang words.

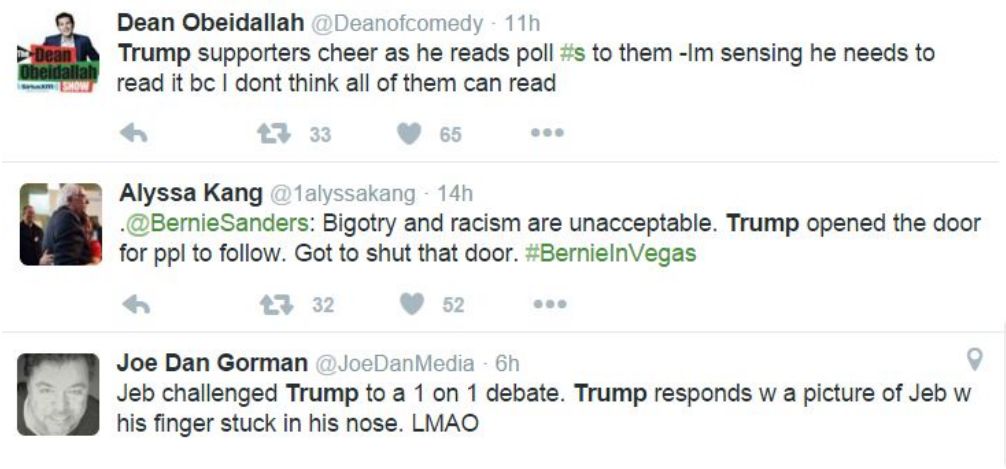

Figure 3: Sample tweets comprising slang words on topic ‘Trump’

“Trump” data set comprises 30981 English tweets only which 2785 of them containing a slang word defined in our slang dictionary.

3. RESULTS AND DISCUSSION

This section contains the network analysis results of the semantic relations that we define under four categories, namely, slang-word co-occurrence, slang-word consecutive co-occurrence, user-slang and location.
Cytoscape, a network representation and visualization tool, is used for analysis [5]. In that framework, we had to re-define our program outputs for creating network and also adding data to our network.

E.g., for our network file (*.sif), we created such relations for slang-co-occurrence.

ok usedWith trump
ok usedWith us
u usedWith donald

Frequencies of the words pairs are represented as the weights of the edges connecting two nodes.

3.1 Slang - Word Co-Occurrence Network

With Stop Words:

Table 1: Top-ten most frequent co-occurring word pairs

word1	word2	freq	word1	word2	freq
u	**rt**	683	**u**	**bald**	309
u	**trump**	596	**u**	**eagle**	309
u	**like**	343	**u**	**attacks**	308
u	**don’t**	336	**u**	**fam**	307
u	**america**	318	**u**	**to**	283

Table 1 illustrates the top-ten most frequent co-occurring word pairs that are dominated by the combinations of different words with slang ‘u’. We can see that most frequent words co-occurring with ‘u’ are consistent with the news in the media. Since Trump has been attacked by a bald eagle, seeing the words “bald, eagle, attacks, trump” are very meaningful [11]. Also, the word “america” most probably comes from Trump’s slogan, which is “Make America Great Again”.

Table 2: Network Analysis parameters for co-occurrence network

Parameter	Value	Parameter	Value
Clustering Coefficient	0.305	**Number of Nodes**	4249
Connected components	1	**Density**	0.001
Diameter	6	**Heterogeneity**	7.8
Radius	3	**Isolated Nodes**	0
Centralization	0.44	**Number of Self Loops**	14
Shortest Paths	100%	**Multi-edge node pairs**	93
Characteristic Path Length	2.9	**Avg. number of neighbors**	5.3

Table 2 shows that for slang-word co-occurrence network, number of connected components are very low (which is 1), meaning that network is strong. Since there are no isolated nodes in the network, it shows that there is no node having closeness centrality 0.

Having nodes with closeness centrality rather than 0 means that information in that network goes very fast. The density of network is very low.

When we look at centralization value, we can see that network is between being evenly distributed or centralized around a node, which doesn't give us enough information to decide. Diameter of the network is 6, showing the longest shortest path in the network. Clustering coefficient of network is not so high, so, if we consider it as "all-my-friends-know-each-other" concept, we can see that co-occurrent words are not highly connected to each other, but connected in average mode.

Table 3: Top-ten words for network analysis of co-occurrence network

Degree Centrality	Betweenness Centrality	Eccentricity	Average shortest path	Closeness Centrality	Clustering Coefficient
u 4 r s ur ok w/ n b bc	u 4 r ur w/ ok s n b bc	qatar leadership credibility weakest thin-skilled slanted performanc e hm setup	post reform imagines dystopian future: breath #numbnutz families living below	u 4 r trump rt to the a of he	democratic hopeful percent caliphate echoes org backing king: joined muslimshe

Table 3 displays the top-ten words for each network analysis metric. When we look at eccentricity results, we can see that, words such as “qatar”, “leadership” can have longer paths in the network, showing that they are mostly used with slang words. Our hypothesis is that this is probably because of the news [10], when Qatar Airways warned Trump that he won’t be welcome in Muslim countries after his speeches against Muslims.

If we look at clustering coefficients, we can see that, words such as “democratic,hopeful, percent, fun, loyal” have lots of neighbors, which shows their co-occurrence strength.

When degree centrality is considered, words such as “u,4,r,s” have more edges rooting from them indicating that they are connected to more words than the others. We can say that these words are the most commonly used slangs amongst the ‘Trump’ data.

If we consider betweenness centrality, we can see parallel results with degree centrality, showing that these words are connecting other subnetworks to each other. Closeness centrality shows how fast a word can go from one node to others, so the words such as trump is the expected result. In addition to these, we see that stopwords are also included in the list, such as “a,to, the” which might be replaced by other words, as we see in the next analysis.

Without Stop Words:

Table 4: Top-ten most frequent co-occurring word pairs

word1	word2	freq	word1	word2	freq
u	**trump**	596	**u**	**fam**	307
u	**america**	318	**u**	**s**	185
u	**bald**	309	**u**	**shut**	150
u	**eagle**	308	**u**	**a**	115
u	**attacks**	307	**s**	**trump**	106

Table 4 illustrates the most frequent co-occurring word pairs with the stop-words removed. When compared to Table 1, we see that the pairs containing words 'rt' (retweet), 'don't', 'like' are no longer in the list but new pairs such as, 'u-s' and 's-trump' are added. ('s' stands for smile or is.) That table is consistent with Table Y, pointing out the eagle attack.

Table 5: Network parameters for slang-word co-occurrence network

Parameter	Value	Parameter	Value
Clustering Coefficient	0.306	**Number of Nodes**	4203
Connected components	1	**Density**	0.001
Diameter	6	**Heterogeneity**	8.1
Radius	3	**Isolated Nodes**	0
Centralization	0.43	**Number of Self Loops**	14
Shortest Paths	100%	**Multi-edge node pairs**	93
Characteristic Path Length	2.9	**Avg. number of neighbors**	4.9

Table 5 indicates that there is no significant difference between co-occurrence network with/without stopwords, except for the decreased number of nodes. When we compare Table 3 with Table 2, we can see that number of nodes in the network is decreased from 4249 to 4203, which provides better performance.

Because both of the networks contain more than 4000 nodes, MCL algorithm failed to produce a result even after an hour later. Therefore cluster analysis for the slang-word co-occurrence networks could not be provided.

Table 6: Top ten words for network analysis of slang-word co-occurrence network (without stopwords)

Degree Centrality	Betweenness Centrality	Eccentricity	Average shortest path	Closeness Centrality	Clustering Coefficient
u	u	office	swallow	u	rnc
4	4	avocado	smiling	4	+
r	r	assoc	#$$$	trump	assholes
s	ur	announces	breath	r	insult
ur	w/	#fox2	#numbnutz	a	remarks
ok	ok	qatar	families	ı	fun
w/	s	leadership	living	for	frm
n	n	credibility	below	ur	loyal
b	b	weakest	$30k/year	donald	posts
bc	bc	thin-skilled	fıred	get	fb

When we compare Table 3 (with stopwords) and Table 6 (without stopwords) , we cannot see any significant difference between the measures of degree and betweenness centralities. However, other results change more significantly.

As we look at eccentricity results, “office, avocado, assoc, announces, fox2” comes in front in the ranking, meaning that they have longer paths in the network and used more with slang words. If we look at the meaning of office, avocado”, we can cite that fact: “A Twitter account called *AvocadoFact* tweeted what seemed like a fun and harmless message on July 1: Who would you rather have in office? rt for an avocado. fav for donald trump?

As of this writing, the avocado is beating Trump, with 61,367 retweets versus Trump's 5,987 favorites.” [9] So, avocado and office have the longest paths in the network.

When the closeness centrality is considered, words such as donald, get are added in the list.

Words corresponding to highest clustering coefficients are changed when the stopwords are removed. Words, such as “assholes, insult, remarks, fun, mc,+” has the highest clustering coefficients, showing that they have lots of neighbors in the network, strongly co-occurring with slang words.

3.2 Slang - Word Consecutive Co-occurrence

With Stop Words:

Table 7: Top-ten most frequent consecutive word pairs (with stop words)

word1	word2	freq	word1	word2	freq
u	**s**	360	**u**	**k**	86
like	**u**	313	**trump**	**v**	84
u	**fam**	307	**v**	**adolf**	82
shut	**u**	146	**n**	**y**	68
the	**u**	98	**fav**	**for**	52

When the most frequent consecutive words are considered, u-s (you smile), like-u (like-you), u-fam (you-family) are the tops pairs (Table 7). fav-for (favourite-for/four) is another consecutive co-occurrence, which suits well for the definition of phrases. Comparison with Table 1 reveals interesting outcomes such that rather than the domination of the word ‘u’, other words such as ‘v’, ‘adolf’, ‘n’, ‘y’ made into the top-list.

Table 8: Network Parameters for Slang - Word Cons.Co. Network (with stop-words)

Parameter	Value	Parameter	Value
Clustering Coefficient	0.034	**Number of Nodes**	1039
Connected components	5	**Density**	0.003
Diameter	8	**Heterogeneity**	3.9
Radius	1	**Isolated Nodes**	0
Centralization	0.23	**Number of Self Loops**	2
Shortest Paths	97%	**Multi-edge node pairs**	65
Characteristic Path Length	3.5	**Avg. number of neighbors**	2.9

Table 8 shows the values for the network parameters. When we consider clustering coefficient, we can see that it is quite low, showing that consecutive co-occurrent words are not highly connected to each other. If network centrality is considered, even distribution of nodes in the network could be observed, having value of 0.23. There is no isolated nodes in the network and network diameter is 8, showing the longest path in the network.

When we compare the consecutive-co-occurrence network with the co-occurrence network, we see that the former is less connected than the first one containing less nodes.

Table 9: Top-ten words for network analysis of slang-word consecutive co-occurrence network

Degree Centrality	Betweenness Centrality	Eccentricity	Average shortest path	Closeness Centrality	Clustering Coefficient
u	gal	money	gal	gal	great
4	4ever	every	4ever	4ever	fool
r	fwd	qatar	into	into	idiots
ur	u	abandoned	fs	fs	winning
s	4	ff	fwd	fwd	kidding
n	r	peeps	thin-skilled	thin-skilled	man
w/	ur	hv	blocked	blocked	next
ok	s	making	fıred	fıred	what's
b	trump	truth	future:	future	let
bc	n		u	u	lol

When we look at eccentricity results in Table 9, we can see that words such as "qatar, money" etc. can have longer paths in the network, showing that they are mostly used with slang words. That's due to the news [10], when Qatar Airways warned Trump that he won't be welcome in Muslim countries after his speeches against Muslims.

If we look at clustering coefficients, we can see that, words such as "great,fool, idiots, winning", have lots of neighbors, which shows their co-occurrence strength.

When degree centrality is considered, words such as u,4,r,ur,s have lots of edges rooting from them,it shows that they are connected to other words more.

If we consider betweenness centrality, we can see that words such as "gal, 4ever, fwd" are highest in ranking, showing that these words are connecting other subnetworks to each other.

Closeness centrality shows how fast a word can go from one node to others, but current words in the list,don't reveal significant results. In addition to these, we see that stopwords are also included in the list, such as "into" which might be replaced by other words, as we see in the next analysis.

Without Stop-Words:

Table 10: Top-ten most frequent consecutive word pairs (without stop words)

word1	word2	freq	word1	word2	freq
u	**s**	360	**trump**	**v**	84
america	**u**	307	**v**	**adolf**	82
u	**fam**	307	**n**	**y**	68
shut	**u**	146	**entering**	**u**	56
u	**k**	86	**fav**	**for**	52

Table 10 shows the top-ten most frequently used consecutive words containing a slang in tweet streams. Comparison with Table 7 shows that the pairs containing stop-words such as ‘like’, ‘the’ replaced with the pairs containing words ‘america’ and ‘entering’.

Table 11: Network Parameters for Slang - Word Cons.Co. Network (without stop-words)

Parameter	Value	Parameter	Value
Clustering Coefficient	0.040	**Number of Nodes**	1202
Connected components	8	**Density**	0.002
Diameter	8	**Heterogeneity**	4.4
Radius	1	**Isolated Nodes**	1
Centralization	0.25	**Number of Self Loops**	2
Shortest Paths	96%	**Multi-edge node pairs**	58
Characteristic Path Length	3.5	**Avg. number of neighbors**	2.7

Table 11 illustrates the networks parameters belong to Slang-Word consecutive co occurrence with stop-words removed. Comparison with Table 8, in which network parameters are displayed for the model still containing stop words, clearly shows that stop-word use affects the parameters. We see that number of nodes increased in the model where stop-words are removed. This change indicates that removing insignificant words allows the consecutive use of infrequent words with the slang, which leds to a variety in the new network.

When we consider clustering coefficient, we can see that it is quite low, showing that consecutive co-occurrent words are not highly connected to each other. If network

centrality is considered, even distribution of nodes in the network could be observed, having value of 0.25. There is 1 isolated node in the network and network diameter is 8, showing the longest path in the network.

Table 12: Top-ten words for network analysis of slang-word consecutive co-occurrence network (without stopwords)

Degree Centrality	Betweenness Centrality	Eccentricity	Average shortest path	Closeness Centrality	Clustering Coefficient
u	cn	avocado	gal	cn	try
4	ff	shows	4ever	fwd	call
r	gal	train	cn	gal	clowns
ur	fwd	leading	fwd	ff	man
s	4ever	tower	ff	4ever	fool
ok	hv	#dumptrump	hv	hv	idiots
n	u	#trumpıstrash	thin-skilled	rants	careful
w/	4	butt	blocked	future:	winning
b	r	truth	fıred	thin-skilled	great
bc	ur	woo	rants	blocked	downs

When we look at eccentricity results in Table 12, we can see that, “avocado, shows, train, tower” etc. such words can have longer paths in the network, showing that they are mostly used with slang words. Avocado issue is again becomes significant in the network (as it is in slang-word co-occurrence network). We also see the words such as “tower”, relating the trump towers.

If we look at clustering coefficients, we can see that, words such as “try,call, clowns, man”, have lots of neighbors, which shows their consecutive co-occurrence strength.

When degree centrality is considered, words such as u,4,r, ur, s have lots of edges rooting from them,it shows that they are connected to other words more.

If we consider betweenness centrality, we can see that words such as “cn, ff, gal, fwd” have the highest centralities, showing that these words are connecting other subnetworks to each other.

Closeness centrality shows how fast a word can go from one node to others, but existing words don’t give significant results.

Clustering Analysis

For consecutive-cooccurrence network model, we compare the clustering results for the models constructed according to with and without stop-word removal. MCL clustering on the network resulted in more than 20 clusters for each network. All of these clusters are formed by at least 5 nodes. (Pdf files illustrating clusters of both network models are provided.)

For both of the network models, most clusters follow a similar pattern where a slang word is in the center connected to its consecutive words (i.e., ur, r, b, bc, ok, w/) . Even though forming similar clusters, we observed that some independent clusters in the model with stop words are connected in the model without stop words. We can conclude that removal of the stop words allows variation in the network by including new and meaningful words to the network. By this way, the clusters which are not connected before because of the noise, can be easily connected.

Among these clusters, some of them present interesting information which will discuss further. For the following discussions, we will focus on the network created without stop-words which provides specificity in the clusters.

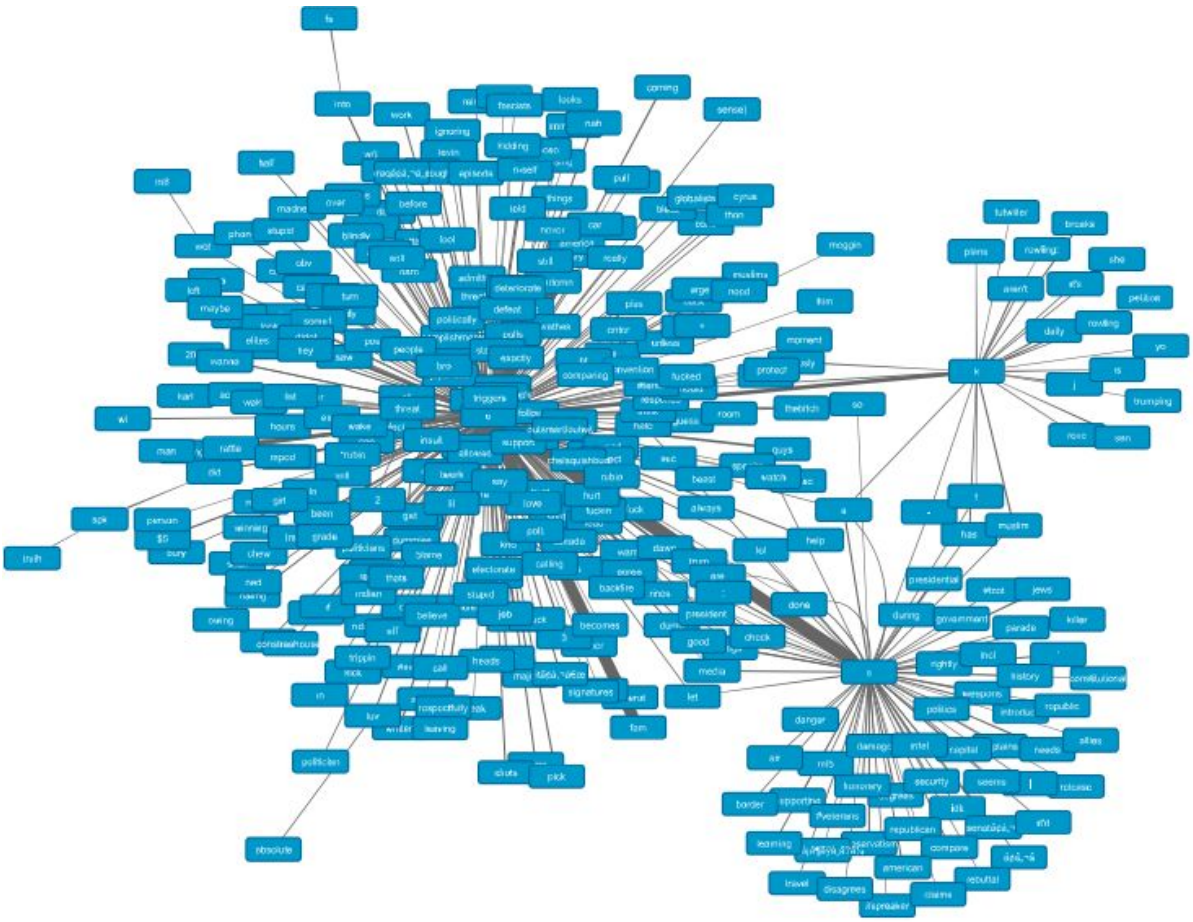

Figure 4: The biggest cluster of the consecutive co-occ. network without the stop-words

In Figure 4, the cluster is formed by the connection of three sub-clusters which are centered around the slang words ‘u’ (you), ‘s’ (smile or is) and ‘k’ (okay). Thickness

of the edges indicate how strong the relationship between two words (i.e. how frequent they are used consecutively in a tweet text).
'k' sub-cluster is connected to 's' sub-cluster via words 'has' and 'muslim','t' and to 'u' sub-cluster directly and through the words 'u', 'should' and 'seriously'. 's' and 'u' sub-clusters are connected directly as well as via the words 'let', 'judge', 'dumb', 'good', 'media', 'president'.

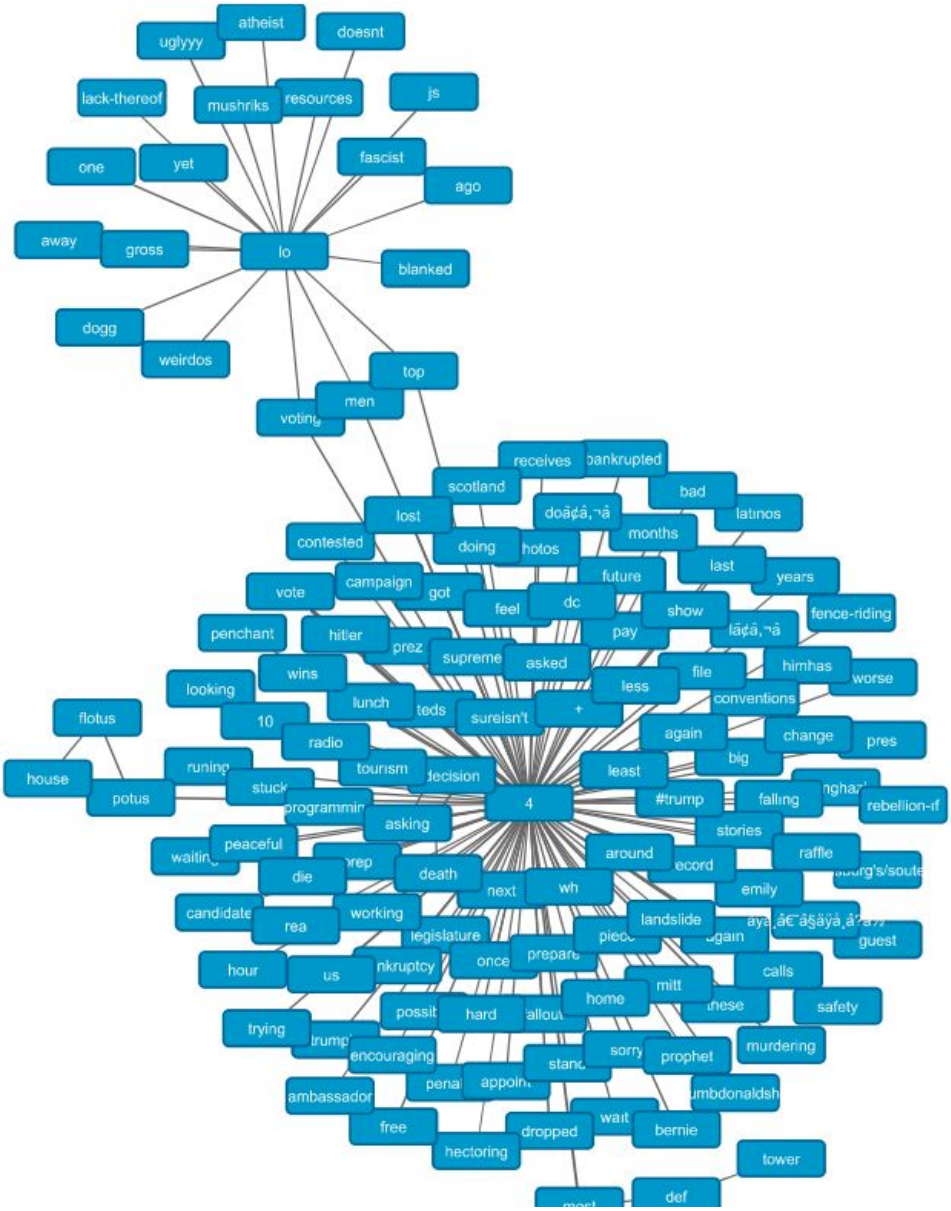

Figure 5:

Figure 5 displays the second biggest cluster of the network, comprising two subcluster centered around '4' (for/ four) and 'lo' (hello). These two subclusters were independent in the model that stop-words are included. But with the removal of the noise, these two sub-clusters connected via the words, 'voting', 'men', and 'top'. The

word gathered around ‘4’ does not follow a visible pattern, whereas the words associated with ‘lo’ are mostly insults such as ‘ugly’, ‘weirdo’, ‘gross’, ‘dog’ as well as some labels like ‘mushriks’, ‘fascist’, ‘atheist’.

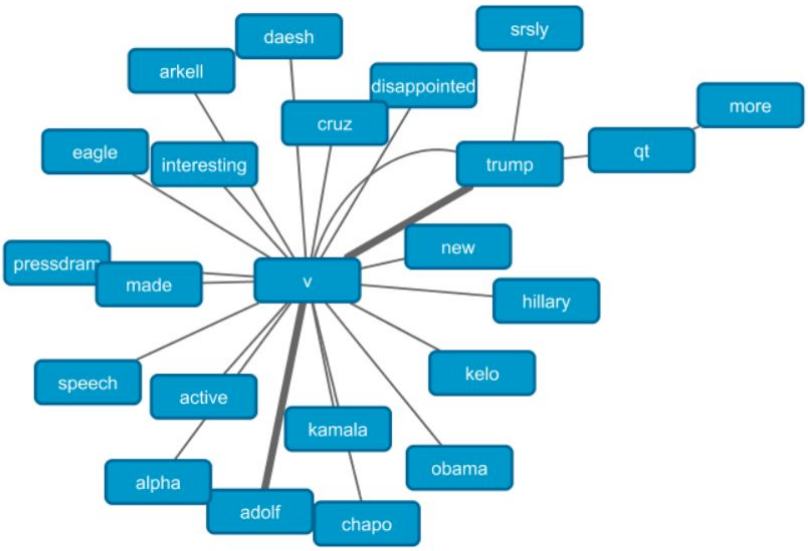

Figure 6: ‘v’ cluster

Figure 6 shows the cluster centered around the slang ‘v’(very). This cluster is the only one that the politicians Obama (Barack Obama), Trump (Donald Trump), Cruz (Ted Cruz) and Hillary (Hillary Clinton) are observed together along with some interesting names such as Adolf (Adolf Hitler), Chapo (Mexican drug kingpin), Daesh (aka ISIS). Quick Twitter search reveals that ‘v’ is mostly used as a substitution for the word ‘versus’.

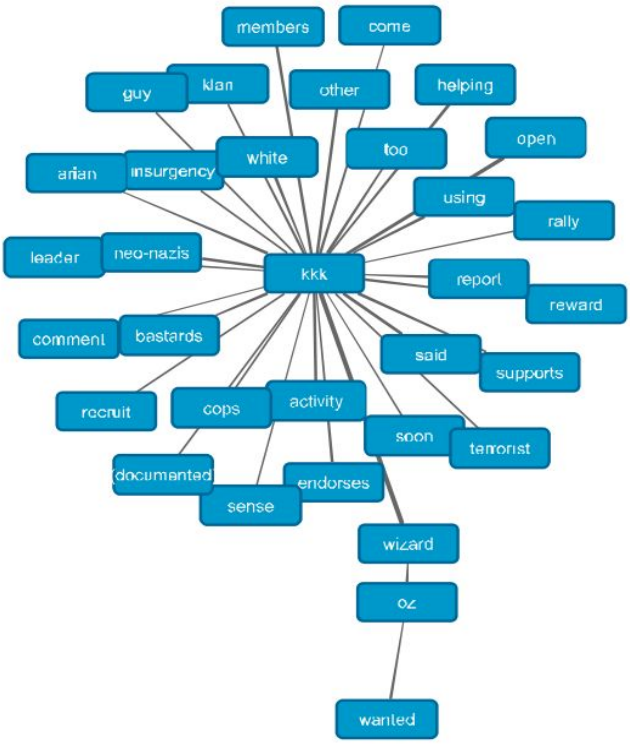

Figure 7: 'kkk' cluster

Figure 7 reveals an interesting cluster with the slang 'kkk' in the center. In our slang dictionary, 'kkk' represents the word 'over', however the other members of the cluster such as neo-nazis, terrorists, cops, white does not make sense with the actual world. A quick research on the keyword suggests that 'kkk' is also used for Ku Klux Klan, a secret fascist organization known with its anti-black demeanors.

3.3 User - Slang

Since this network consists of the nodes connecting the user and the slang words that they use, stop-word removal process has no effect on the network parameters.

Table 13: Network Parameters for User - Slang Network

Parameter	**Value**	**Parameter**	**Value**
Clustering Coefficient	0.0	**Number of Nodes**	2136
Connected components	19	**Density**	0.001
Diameter	12	**Heterogeneity**	9.0
Radius	1	**Isolated Nodes**	0
Centralization	0.43	**Number of Self Loops**	0

Shortest Paths	94%	**Multi-edge node pairs**	0
Characteristic Path Length	4.1	**Avg. number of neighbors**	2.3

When we look at the values of the user-slang network in Table 13, we can see that clustering coefficient is 0,which shows that all the nodes have neighbors less than two.
If network centrality is considered, the value doesn't reveal significant information. There is no isolated nodes in the network and network diameter is 12, showing the longest path in the network. Connected components are 19, showing that network is weak.

Table 14: Top-ten most-frequent words for user-slang network

Degree Centrality	Betweennes s Centrality	Eccentricity	Average shortest path	Closeness Centrality	Clustering Coefficient
u s 4 r v n ok ur kkk bc	huh hay ff gal w/o flotus 4ever spk ova def	rt for sportacu _gwen_e _madz_ind 1975cayl 420zqua ASEXUALEV avidalessi BrigidMaguir cappvccin Chandler_Kruc	japolok KloppKloppKlop L_Buchfin LvMyLoca mc_money TigerTownBab hahaha rt for sportacu _gwen_e _madz_ind	ff flotus 4ever huh hay ashley_simpki n srsly spk Blogserve fwd	0

If we look at clustering coefficients in Table 14, we can see that it is 0, which shows that all the nodes have neighbors less than two.
When degree centrality is considered, words such as u,s,4, r, v have lots of edges rooting from them,it shows that they are connected to other words more.
If we consider betweenness centrality, we can see that words such as "huh, hay, ff, gal" have the highest centralities, showing that these words are connecting other subnetworks to each other.
Closeness centrality shows how fast a word can go from one node to others, words such as "ff, flotus, 4ever, huh" are going in a fast way in the network.

3.4 Location - Slang

In this network model, instead of the usernames, the location of the users are connected with the slang words they use. Therefore, only change in the model is the type of the nodes and similarly, stop-word removal process has no effect on the network parameters which are shown in Table 15.

Table 15: Network parameters for location-slang word network

Parameter	**Value**	**Parameter**	**Value**
Clustering Coefficient	0.004	**Number of Nodes**	1062
Connected components	8	**Density**	0.003
Diameter	7	**Heterogeneity**	5.9
Radius	1	**Isolated Nodes**	2
Centralization	0.4	**Number of Self Loops**	0
Shortest Paths	97%	**Multi-edge node pairs**	0
Characteristic Path Length	3.3	**Avg. number of neighbors**	2.7

When we consider clustering coefficient, we can see that it is quite low, showing that location-slang words are not highly connected to each other. If network centrality is considered, even and a little centralized distribution of nodes in the network could be observed, having value of 0.4. There are 2 isolated nodes in the network, showing that there are some nodes in the network having closeness centrality as 0. Network diameter is 7, showing the longest path in the network.

Table 16: Top-ten words for network analysis of location-slang network

Degree Centrality	Betweenness Centrality	Eccentricity	Average shortest path	Closeness Centrality	Clustering Coefficient
u s 4 ok empty_field r n v k	gal spk 4ever u empty_field 4 ok v s	Vancouver B sux Sextown U West-Central Afric ho Colorado, US Iowa, US Pittsburgh, P Houston, Texa	Vancouver B Sextown U West-Central Afric sux U.S Sarasota, F Ky. New Concor Charlotte, N Macomb, Michiga	spk Kent - England qt gal Fall Creek, W oz 4ever England North Britai	World laguna beach Rhode Island, US Ft. Worth T Paris leicester u Florid New Jersey, US N

kkk	n	Saco, Main	Riverside, C	Conch Republic & Blue Ridg	f

When we look at eccentricity results in Table 16, we can see that, “Vancouver, Sextown, West-Central Africa, Colorado” etc. such words can have longer paths in the network, showing that slang words are used from these locations mostly. If we look at clustering coefficients, we can see that, words such as “World,laguna beach, Rhode Island, Ft. Worth T, Paris”, have lots of neighbors, which shows their location-slang combination strength.
When degree centrality is considered, words such as “u,s,4, ok” have lots of edges rooting from them,it shows that they are connected to other words more.
If we consider betweenness centrality, we can see that words such as “spk, Kent-England, qt, gal” have the highest centralities, showing that these words are connecting other subnetworks to each other more.

Clustering analysis for the Location-Slang network does not provide a useful insight about the network because the majority of the locations are empty, which we refer as *empty_field*. The biggest cluster, which comprises bc, plz, v, lo, kkk, 4, m, c, dem, r, ur subclusters, is formed because of the empty_field node.

4. CONCLUSION

With this study we first provide a slang ontology for Twitter that defines relations between the actual words and the slang words as well as the relations between the users and the words. Then based on these pre-defined semantic relations, we process the Twitter data and provide network analysis for slang-word co-occurrence, slang-word consecutive co-occurrence, user-slang, and location-slang network models.
For case study, we collected data centered around the keyword ‘Trump’ (i.e., Donald Trump) which also links with 2016 US elections.

The network analysis results show that investigation of co-occurrence of the slangs with the other words that are used together reveals a general picture about the topic of interest and the current debate such as bald eagle example for the Trump data. Consecutive co-occurrence network model, on the other hand, provides patterns for the frequently used slang words. For example, we see that ‘v’ keyword is frequently used in Trump data mostly to compare a politician with Trump.

There are two main limitations of our work first of which is the existence of a narrow slang dictionary containing almost 250 words. Therefore, a future work based on an extended slang dictionary will increase the number of tweets that are processed. The

second limitation arises from the lack of semantic integrity of the locations such as the case where locations UK and Liverpool (a city in UK) are treated as two different places. In an ideal model semantic mapping of Liverpool to UK will provide much more information for the network analysis. Besides, parsing problems such as separating 'St Petersburg' into two words and use of the user-generated location information with misuse (i.e., home, idk, neverland) creates a noise in the processed data.

ACKNOWLEDGEMENT

We sincerely thank Ş. Betül Bilgin who contributed the design of the idea, construction of the ontology and early-development of the data processing.

REFERENCES

1. Lee, Deirdre, and Mohammad Waqar Adegboyega Ojo. "Utilising Linked Social Media Data for Tracking Public Policy and Services."

2. Clark E., Araki K (2011) . Text normalization in social media: progress, problems and applications for a pre-processing system of casual English, Procedia - Social and Behavioral Sciences 27 (2011) 2 – 11, Elsevier Publishing

3. http://med.bioinf.mpi-inf.mpg.de/netanalyzer/help/2.6.1/#cksDist

4. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*. 2002;30(7):1575-1584.

5. Shannon, Paul, et al. "Cytoscape: a software environment for integrated models of biomolecular interaction networks." *Genome research* 13.11 (2003): 2498-2504.

6. https://en.wikipedia.org/wiki/Network_science#Diameter_of_a_network

7. http://www.netlingo.com/acronyms.php

8. https://jena.apache.org/

9. http://mashable.com/2015/07/05/donald-trump-vs-avocado/#cwafAcKazgqw

10. http://money.cnn.com/2015/12/10/news/companies/donald-trump-qatar-airways-ceo/

11. http://www.buzzfeed.com/stephaniemcneal/bald-eagle-vs-balding-man#.vbyE4k0e3