Network and Complex Systems
ISSN 2224-610X (Paper) ISSN 2225-0603 (Online)
Vol.5, No.4, 2015

Knowledge Discovery and Complex Network Dynamics in Social Media Space

Edward Yellakuor Baagyere[1,2] Zhen Qin[1*] Xiong Hu[1] Qin Zhiguang[1]
1.School of Information and Software Engineering, University of Electronic Science and Technology of China, No. 4, Sec. 2, Jianshe North Road, Chengdu, Sichuan Province, China
2.Department of Computer Science, University for Development Studies, Box 1350TL, Tamale, Ghana
* E-mail of the corresponding author: qinzhen@uestc.cn;baagyere.uestc@gmail.com

This work is supported in part by the National Basic Research Program of China (No. 2013CB329103), the National Science Foundation of China (No. 61133016, 61300191 and 61370026), the Ministry of Education-China Mobile Research Foundation (No. MCM20121041), the Fundamental Research Funds for the Central Universities (No. ZYGX2013J003, ZYGX2013J067, ZYGX2013J073and ZYGX2013J083))

Abstract
Pattern discovery and correlation in text data have been research hotbed in recent times. However, a composite model that captures patterns and correlations as a quantitative measure in social media space is yet to receive much research attention. The paper therefore analyzed social media data from Twitter about the 2014-FIFA World Cup both as lexical text and a complex network system. Quantitatively it is discovered that the 140 character upper bound in Twitter does not have negative impact on the formation of ideas. For as a lexical text, the following key statistics were confirmed: the distribution of the words in the corpus obeys a Zipf's law, 3-character length words accounted for almost 22% of the corpus and the distribution of the article *"the"* also follows a Zipf's or power-law. Moreover, the three most frequent terms related to the world cup event, that is (*url, worldcup, rt*) account for about 14.5% of the corpus.
In particular, the corpus is modeled as a network, $G = (V, E)$ where V is the set of vocabularies in the corpus and E is the set of bigrams (two words phrases). An algorithm is developed and implemented in python to obtain the bigrams from the corpus. Using concepts from graph theory, the bigram network is analyzed and the results show compelling facts about text network. Firstly, all the characteristics of complex networks known in literature are observed in the bigram network. These include the degree distribution, which is observed to follow power-law with degree exponent γ value of 2.14. Secondly, the average path length of words is observed to be 4.78, which is within the "small world" categories. Thirdly, other complex network characteristics such as eigenvector and betweenness centralities metrics are observed within the bigram network both having weak power-law distributions as observed in other complex networks in literature.
These findings call for the need to study the topological characteristics of text data and comparing their structural properties to that of known complex network metrics in literature. The results will be of great importance in studying complex systems. Also the application areas of these findings are numerous ranging from information retrieval, data compression to information security.
To the best of our knowledge, this is the first work that studied the textual and topological structure of text from social media platform as a complex network and analyzed important topological properties of complex network on it.
Keywords: complex network, bigram, media space, Twitter, information science

1. Introduction

The social media space is an open natural laboratory that contains a lot of information that can be harnessed for many research purposes. Social media refers to a set of online tools that is built on the ideas and technology of Web 2.0 with the main objective of creating and exchanging content between users [1]. These tools are of different forms and are for different purposes. One of such social media tools is Twitter. Twitter is a multi - dialogue *microblog* tool that allows users to post short status updates, called tweets, that appear on the user timelines. The tweets are limited to 140 characters length, earning Twitter the name *microblog* social media platform and these tweets can include various entities, and geographical locations. As such, people can share a lot of short messages instantly on almost all digital devices making Twitter a good social media tool to discuss noteworthy events that have both global and local significance. These messages together with the metadata about the authors, provide never before, a whole library of data that can be analyzed to understand a wide variety of issues ranging from worldwide events such as the FIFA World cup to socio-political issues such as political polarization to geographic and demographic lexical variations.

Text mining and information retrieval have also become an important research area in recent times.

Text mining is the process of discovering the hidden patterns and relationships in text data. The basic aim of text mining is for discovering of patterns and relationships within text data[2,3]. These patterns and relationships discovering have many application areas; these could include text classification, text clustering, ontology and taxonomy creation, document summarization and latent corpus analysis. It is proven that if you are able to know the words with the highest frequency in an English text, you can be able to know most of the terms in the given English text [4]. Also with the ability to process large data sets we can answer many interesting questions ranging from linguistics to complex network theory. For example, J.B. Michel *et al*. [5] processed 15% of the digitized google books content making up of about 4% of books ever printed, in order to investigate the diffusion of regular English verbs and so as to perform a time series analysis on a person's cultural influence. M.E.J. Newman [6]observed that the cumulative distribution of the number of words occurrence in a typical piece of English text (Moby Dick, a novel by Herman Melville) follows a power law. All these and many others point to the importance of text mining not only as an academic activity but as a worthwhile course for society at large.

Most of these researchers essentially looked at the morphological properties of text data. However, their research were constrained by data availability in diverse forms authored by several millions of people on the same subject matter[3,4,5,6,7]. Also, the inter-relationship of words as a complex network have not received enough research in the domain of social media. To this end, this paper investigates the corpus complexity of the 2014 - FIFA World Cup (FWC) in order to analyze the distribution and the inter-relationship of words in it and thereby made the following contributions:

- A detail statistical analysis on social media data is outlined. The statistical results show that the 140 characters length upper bound on tweets have no negative effect on opinions formation in tweet as most of the results are in agreement with what is known in literature, such as the Zipf's law, e.t.c.
- We discovered the dominance of three words within the corpus as result of the 140 characters upper bound within Twitter.
- We developed a bigram algorithm and used it to model a complex network in order to study complex network properties on text data.
- We further confirmed that most of complex network properties such as degree distribution, shortest path length, eigenvector centrality measure, e.t.c. in the bigram network are within the range of most social networks and "small world networks".
- We established that complex network tools can distil messy data into patterns and inter-relationships. These findings have relevance in many other fields such as information security, knowledge discovery, information retrieval, e.t.c.

The remaining Sections of the paper are organized as follows: Section 2 discussed various methods employed in getting the users' tweets, how the data was cleaned and how statistical measures are used to analyzed the tweets. In Section 3, the topological characteristics of complex networks are studied on the bigram network generated. Section 4 gave some further analysis and applications of the research findings. Related works in literature are studied in Section 5. Conclusion and future work are shown in Section 6.

2. Methodology

This section describes how the tweets were collected and cleaned to form the corpus for the analysis.
Lexical and Social Network analysis tools and techniques are then applied on the cleaned corpus to map out the hidden patterns in it.

2.1 Data Collection and Processing

The corpus for this paper was obtained by streaming the Twitter Application Programming Interface (API) to sample public data from the Twitter firehose during the 2014 FIFA World Cup period. The World Cup began on 12th June with a group stage and concluded on 13th July with the championship [8]. Twitter makes up 1% of all tweets available in real time through a random sampling technique that represents the larger population of tweets and exposes these tweets through the Streaming API[9]. Streaming the API therefore provides a way to sample from worldwide information in as close to real time as possible.

The sampling of the tweets was done during the peaks time of the World Cup. The first sample was taken 30 minutes before the start of a match, the second sample taken during the 90 minutes play time and the last sample taken within 30 minutes after a match. Therefore making access to 1% of all public tweets during the peak periods of the World Cup is very significant to do any meaningful analysis.

After harvesting the tweets for the 3 weeks period, the tweets were cleaned using a Python Script to get the necessary data for the analysis which is termed as corpus. The corpus contains messages written by users in different languages from all over the world which may contain hashtags, user mentions and other characters such as emotion icons, etc.

2.2 Statistical measurements

Text conveys both quantitative and qualitative information; it records and communicates events, data, facts, and opinions. Therefore, statistical measurement of text is relevant to understanding these textual properties and also the inter-relationship among words in text data. The corpus complexity in social media space thus affords us the opportunity to explore tweets for statistical properties and relationship in the Twitter data. The following sections outlined some statistical properties and graphical presentations of text data.

2.2.1 Text Statistics

Understanding the statistical nature of text is fundamental to building efficient models for information retrieval. The statistical models of word occurrences for instance are used in many core components of search engines, such as ranking algorithms, query transformation, and information indexing techniques.

It has been observed that the statistical distribution of word frequencies in a text is very skewed. There are few words that have very high frequencies and also there are many words that have low frequencies. This phenomena is often described by the Zipf's law, which states that the frequency of the rth most common word is inversely proportional to r, or alternatively, the rank of a word times its frequency (f) is approximately a constant (k) [10]. This is mathematically expressed as:

$$r.f = k \quad (1)$$

Which in turn is expressed in a probability form as:

$$rP(r) = c \quad (2)$$

where $P(r)$ is the probability of occurrence of the rth ranked word, and c is a constant. The value of c in English text is $\cong 0.1$.

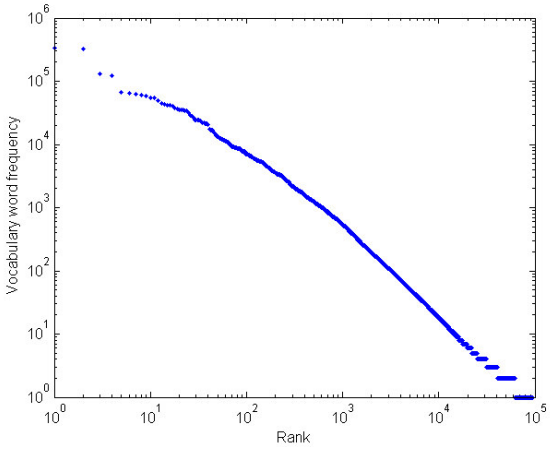


Figure1: FWCC Words distribution

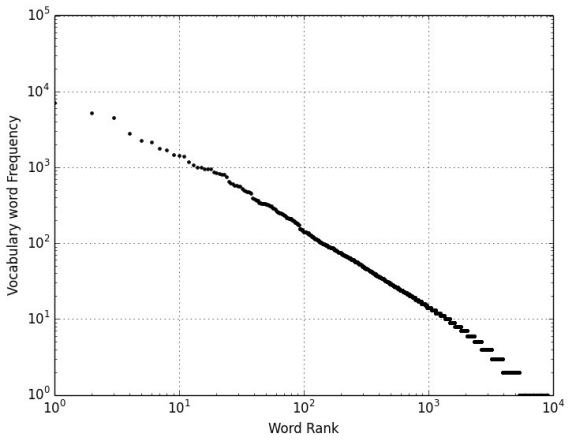


Figure 2: US-IAC f words distribution

2.2.2 Words Distribution in Corpus

The FIFA World Cup Corpus (FWCC) has 5,398,183 words with only 93,700 unique vocabulary collections, which is about 57 times smaller than the total number of words. Thus, the over 5 million words in the corpus is made up by the repetition of the 94 thousand unique words. For the purpose of comparative analysis, the FWCC is compared with that of the US-Inaugural Address Corpus (US-IAC) which is made up of a collection of the presidential speeches of all United State Presidents from 1789 to 2013. The US-IAC is made up of 134,238 words and 9,041 vocabulary set. One key feature about these two corpora is that the distribution of the vocabulary sets are not evenly distributed. Figures 1 and 2 show that the distribution is that of a Zipf's law (power law). The curves approach a linear behavior at the middle ranks and deviate from a linear behavior at the extreme ends. This linear behavior in a logarithmic space is also called power law distribution which is a characteristic behavior of most languages [6,10] and many real-networks such as scientific collaboration [11], the Web [12], wealth distribution of nations [6]. A further observation of Figure 3 showed that there are four words at the extreme lower rank of the distribution that are standing alone. These words are *url, worldcup, rt* and *the*. The *url* stands for the various Uniform Resource Locator addresses in the corpus, and *rt* stands for all the re-tweeted messages. The first three of these words are keywords that have a direct relation with the corpus, while *"the"* is a word that is known to be the most frequent word in every English text or corpus and this is confirmed in the US-IAC in Figure 4. It can therefore be inferred, in the case of the FWCC, that majority of the text was written in English language of which most were re-tweeted messages, and there are a lot of *url* addresses within

the corpus.

The statistical distributions of the unique vocabulary set in the FWCC is shown in Table 1. It can be observed that there is a huge disproportion in the way these vocabularies are distributed within the corpus. For less than 3% of the vocabularies account for almost 88% of the words in the corpus.

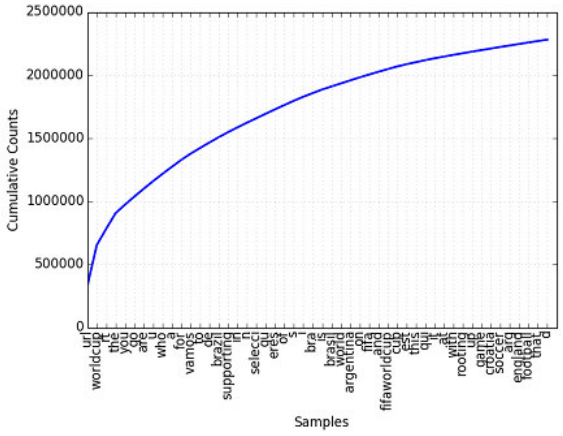


Figure 3: A Sample of FWCC distribution

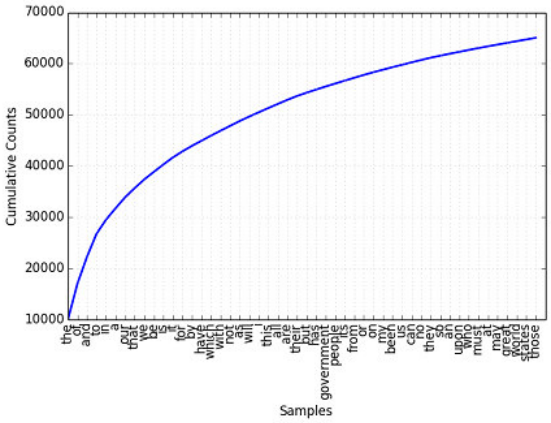


Figure 4: A Sample of US - IAC distribution

Table 1: Statistics of the most frequent words in the FWCC within different ranks

Rank Interval	1-7	1-42	1-249	1-1147	1-2254
Percentage of Vocabulary	0.01%	0.04%	0.27%	2.45%	2.45%
Total Number of Occurrence	1,099,340	2,260,732	3,514,807	4,447,485	4,744,868
Percentage of whole corpus	20.37%	41.88%	65.11%	82.39%	87.90%

2.2.3 Lexical Diversity of corpus

Lexical diversity is an important statistic that quantifies the richness of a text by the distribution of the vocabulary set within it. Lexical diversity is defined as the ratio of unique words to the number of total words in a corpus. For example, a lexical diversity of 1.0 would mean that all words in a given corpus were used uniquely whereas a value of 0.0 implies more duplicate words in a corpus.

In a social media space like Twitter, lexical diversity might be interpreted in different ways. In the context where multiple authors are talking about the same topic such the World Cup even in this case, a much lower than expected lexical diversity might also imply that there are a lot of "group thinking" going on or there is a lot re-tweeting going on, in which the same information is been passed on from one author to the other. Thus, the lexical diversity of the FWCC under study is 0.0175, which means that about 1.74% of the vocabulary set were uniquely used in the corpus. The low lexical diversity maybe due to the high re-tweeting nature of the corpus. The high re-tweeting phenomena can be deduced from Figure 3, where *rt* is the third ranked term in the corpus, thus accounting for the low diversity of opinions in the corpus.

However, for comparative analysis, the lexical diversity of the Inaugural Address Corpus is 0.067, which means about 6.7% of the vocabulary set were unique in the text, that is there are more "individuality thinking" within this corpus as compare to that of the FWCC.

2.2.4 Power law in word intermittency

An important property of language is the words intermittency or burstsiness within its structure. This property asserts that the occurrence of a word in a given text has the tendency to repeat itself in a specific pattern that resemble that of bursts occurrences. In the FWCC, the third highest ranked word is the determiner "*the*". The average number of words interval within which "*the*" appeared in the corpus is about 43 words, with a median value of 25 words.

That means, after every 43 words sequence, one has a high chance of encountering the word "*the*" in the corpus. But the distribution intervals of "*the*" is not even, as the mean distribution value is statistically higher than the median value. Figure 5 highlights this scenario. The topmost part of the figure shows the burstsiness behavior of *"the"* with the black line marking out the mean interval length. The uneven distribution of the interval lengths is so obvious, many of the interval lengths are concentrated between 1 and 43 while majority of the interval lengths are very widely distributed between 43 and 3000, very far above the mean value of 43. The

lower panel of the diagram shows the histogram distribution in linear space and in logarithmic space of the interval length distributions of the consecutive occurrences of "*the*" in the corpus. The consecutive occurrences of "*the*" in the FWCC is confirmed to follow a power law distribution as shown by the logarithmic space distribution portraying almost a straight line with a negative slope.

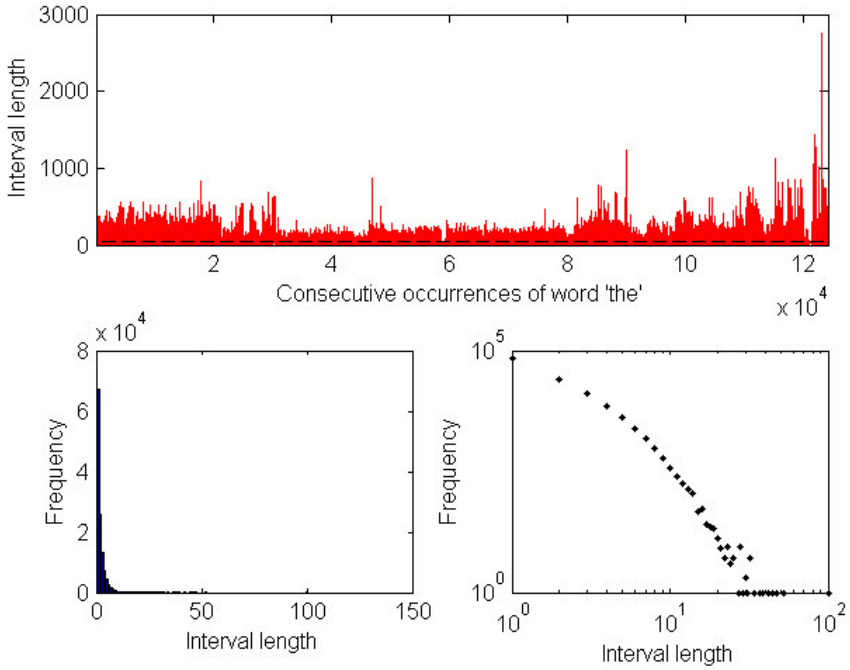


Figure 5: Interval lengths between consecutive occurrences of word "*the*" and their corresponding histogram, in linear space (bottom-left panel) and in logarithmic space (bottom right panel)

2.2.4 Words Length Distribution

The word length distributions of the FWCC is in line with what Zipf posited, that most frequently used words in a language are in average the shortest ones [10]. This observation can be verified directly from Figure 6 and from Tables 2 and 3. From the two Tables, it can be observed that about 70% of the FWCC corpus are within 2 to 8 words length while 80% of the US-IAC corpus are within 2 to 7 words length. Also from Figure 6, it can be observed that the average word length increases alone the x-axis showing very few words with high average lengths, as we move in the frequency rank of words. The longest word is 120 character long and is only one word. This phenomena shows the natural auto-compressing property inherent within languages which is similar to most compression algorithms such as the Hoffman compression algorithm. These statistics are very interesting as they provide a reasonable starting point in understanding what the 140 character content of users tweets have revealed about the 2014 FIFA World cup. These statistics can further be exploited and applied in the design and analysis of future information dissemination systems, as a road map in forecasting future events and can even aid in the design of current social networking systems, just to mention but few of them.

Table 2: Statistics of the most frequent words length in the FWCC

Word length Rank	3	2	4	8	5	6
% of occurrence in corpus	21.81%	12.45%	10.22%	9.49%	7.67%	6.22%

Table 3: Statistics of the most frequent words length in the US-IAC

Word length Rank	3	2	4	5	6	7
% of occurrence in corpus	21.50%	20.11%	13.76%	9.76%	8.02%	7.46%

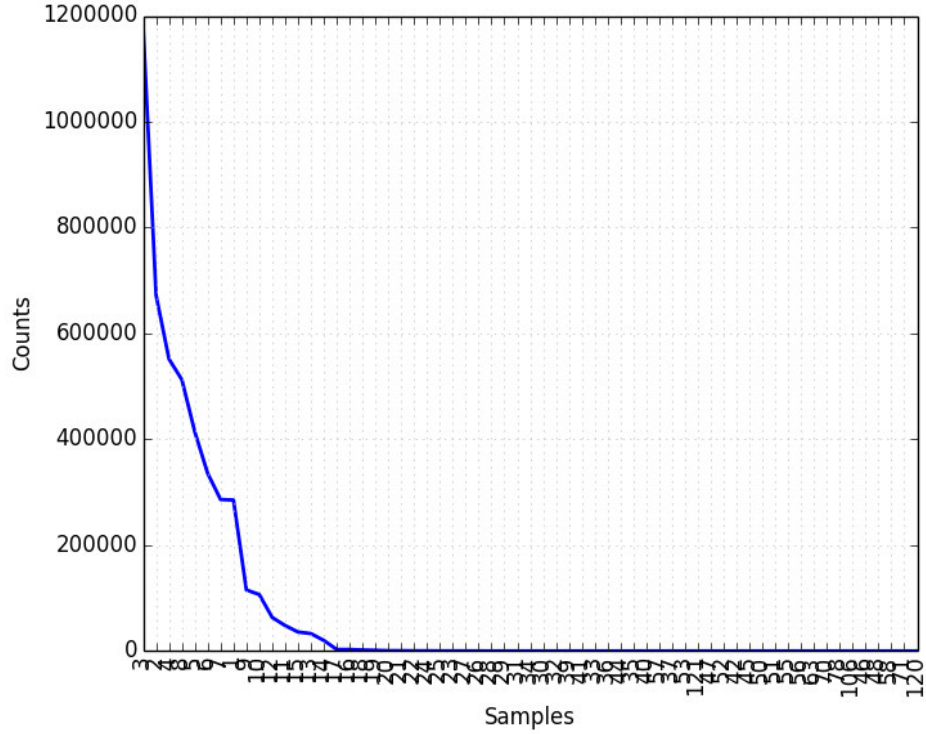


Figure 6: Words Length Distribution

3. Bigram complex network

Text can be seen as a complex network of words, where these words link with each other to form sentences and sentences linked with each other to form paragraphs which in turn link up to form a complex text or idea. A network can therefore be generated using a bigram of words which can enable us to known which two words phrases co-occur often, and which do not. These can offer a blue print for information retrieval, algorithm optimization, information encoding, information security and many others.

To this end, we developed a bigram algorithm and then wrote python codes to generate bigram of words in order to build a bigram network. The bigram network is modeled as a graph $G = (V, E)$, where V is the set of vocabulary set and E is the set of bigrams. The algorithm for generating the bigram network is outline below.

The network generated using Algorithm 3.1 is named FWCC graph or network. The FWCC graph is a directed graph that consists of 250,528 vertices and 481,306 edges. The structural properties of the network are outlined in the following sections.

Algorithm 3.1: BIGRAMGRAPH(W, V, E)

$W \leftarrow WordsInAGivenCorpus$
$V \leftarrow VocabularyOfWordsInCorpus$
$E \leftarrow BigramOfWordsInTheCorpus$
while $v_i \in V$
do { **if** $e_i \in E$
then $G \leftarrow (V, E)$
return (G)

3.1 Structural Properties of the Bigram network

A complex network contains some key properties that when measured and analyzed help us to understand the nature of the complex network and the underlying factors that brought the network into existence. The common structural properties of complex networks are measured and analyzed in this section. Table 4 shows the statistics of some of these structural properties on the FWCC network. For example, from the Table, it shows that the

FWCC network is about 90.4% connected, have an average degree of 1.92 and density of 1.534e−05. The FWCC network is also very modular in structure with over 11,384 communities. These statistics therefore give us a brief idea about the structure of the FWCC network. Another interesting structural property of a complex network is it degree distribution. The next section outlines the degree distribution of the FWCC network.

Table 4: Network Statistics

Statistical Properties of Bigram Network								
Network	V	L	$< k >$	DAC	λ_1	#Cliques	$C_{c\,size}$	MC
WC-Bigram	250528	481306	1.92	-0.09	113.04	11315	90.4%	0.531

(5)

3.1.1 Degree distribution

A complex system can easily be understood if its constituents are mapped out. The constituents underlying complex networks are the nodes and edges that interconnect with each other. A key property of a node in a complex network is its degree (k), which represents the number of links it has to other nodes. Thus the larger the node's degree, the "more important" the node is in a network. The degree distribution $Pr(k)$ of a network is obtained from the nodes degrees and is defined as the probability that a randomly selected node within the network has degree k[15].

The degree sequence of a graph G is obtained by listing its respective nodes degrees. When the node degrees of a complex network are considered in aggregate, through the degree sequence $k_1, k_2, k_3, \ldots, k_n$ various important structural properties of the network can be studied. For example the average nodes degree $< k >$ for undirected network can be obtained from the degree sequence as:

$$< k > \equiv \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2L}{N} \quad (3)$$

and for a directed network, $< k > = L/N$ (4)

where L is the total number of links in the network. The degree sequence and distribution have relevant data to the understanding of the structure of a network. For example, in the studies of empirical networks, the skewness of the degree sequence tells the presence or absence of hubs in the network, which in turn informs the resilient of the network to attack and to the spread of disease [13, 14].

The degree distribution of the FWCC network is shown in Figure 8(a) and is skewed to the right. Such networks are called power law networks and are described in the form $Pr(k) \sim k^{-\gamma}$ where γ is the degree exponent. Because power-law networks are free of any characteristic scale, such networks with a power-law distribution are called scale free networks. Thus the FWCC network degree distribution is best fit with a power law distribution with a γ value of 2.14 and *xmin* value of 22 using the Python package for analysis of heavy-tailed distributions [16]. For the sake of comparison, the FWCC network is simulated as a random network with the same number of edges and vertices. This network is named FWCCR. The degree distribution of FWCCR is shown in 8(b). The two networks though having the same average degrees, number of vertices and edges, are sharply different in their degree distributions. This shows that the words relationship in the FWCC was not randomly distributed but created by a power-law property, where few hubs are having so many connections as high as 12,207 which absent in the case of the FWCCR network.

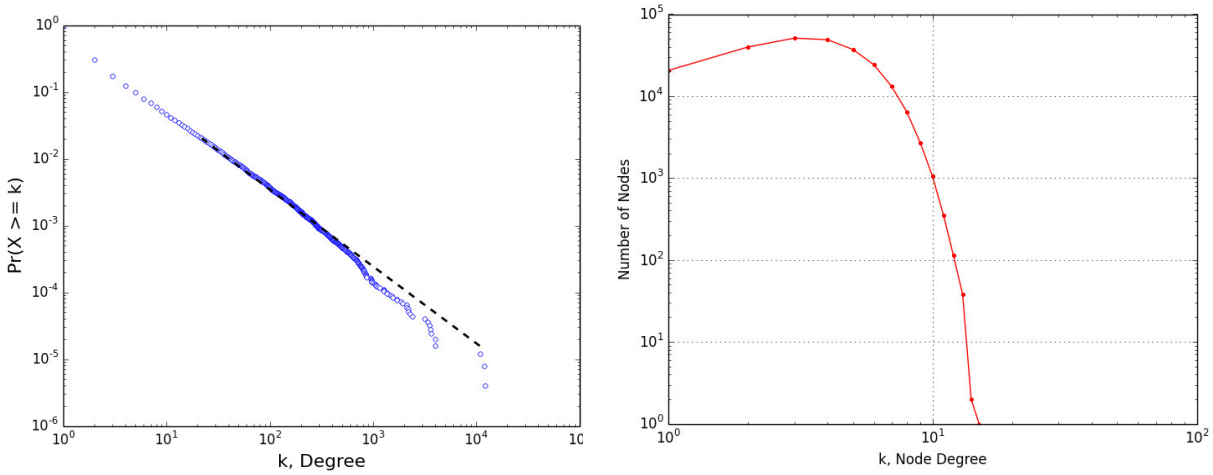


Figure 8: Nodes Degree Distributions

3.1.2 Joint Degree Distribution

The degree distribution alone is not representative enough in describing a network topology because there can be more than one network with the same degree distribution. Also it does not give us information about which node is connected to which node within the network. The Joint Degree Distribution (JDD) is a network measure that attempts to measure how nodes are connected with each other in a network. The JDD is mostly approximated by the degree correlation function k_{nn}, the k nearest neighbor average node degree, in large networks. An increasing k_{nn} indicates the tendency of higher-degree nodes to connect to other high-degree nodes; a decreasing k_{nn} represents the opposite trend. The Pearson degree correlation measure is the most condensed way to characterize the degree - link structure of a network. A negative value of the Pearson correlation indicates that nodes of dissimilar degree tend to be linked and positive value indicate otherwise [17]. The k_{nn} plot for the FWCC network is shown in Figure 9(a) with the Pearson degree correlation value inserted. The Figure and the Pearson correlation value show that the FWCC network has high degree nodes connecting to low degree nodes.

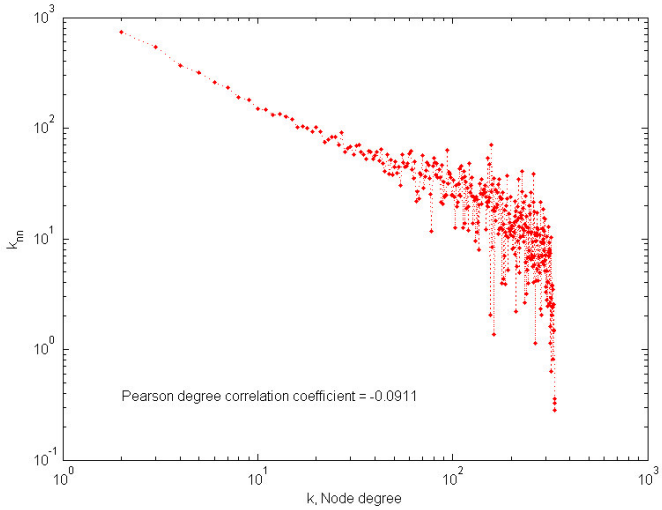


Figure 9: Nodes Nearest Neighbors Distributions

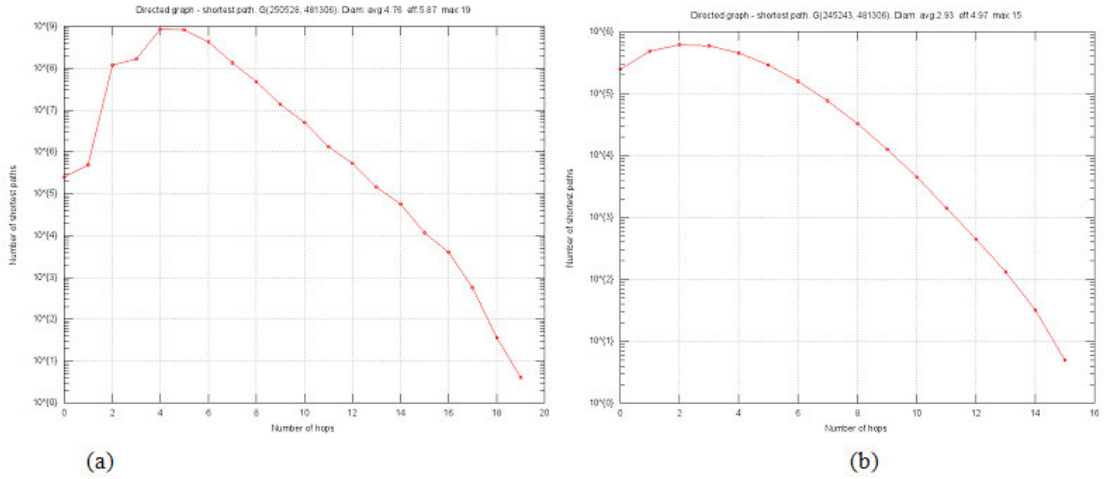
3.1.4 Network distance statistics

Another key characteristic of complex networks is the distance statistics, how far apart are the nodes in the network. The distance between two vertices v_i and v_j in a complex network is expressed as the shortest path between them. The distance statistics for the FWCC graph and FWCCR are shown in Table 5. Also Figures 10(a) and 10(b) show the distance distributions for FWCC and FWCCR respectively. The two Figures and the Table show that the distance between nodes (words) in the networks are very small as the average distances scaled

logarithmically with the network size thereby exhibiting a small world property.

Table 5: Distance Statistics for FWCC and FWCCR networks

Network Type	90% effective diameter	Average Path Length	Diameter
FWCC	5.87	4.78	19
FWCCR	4.97	2.93	15



(a) (b)

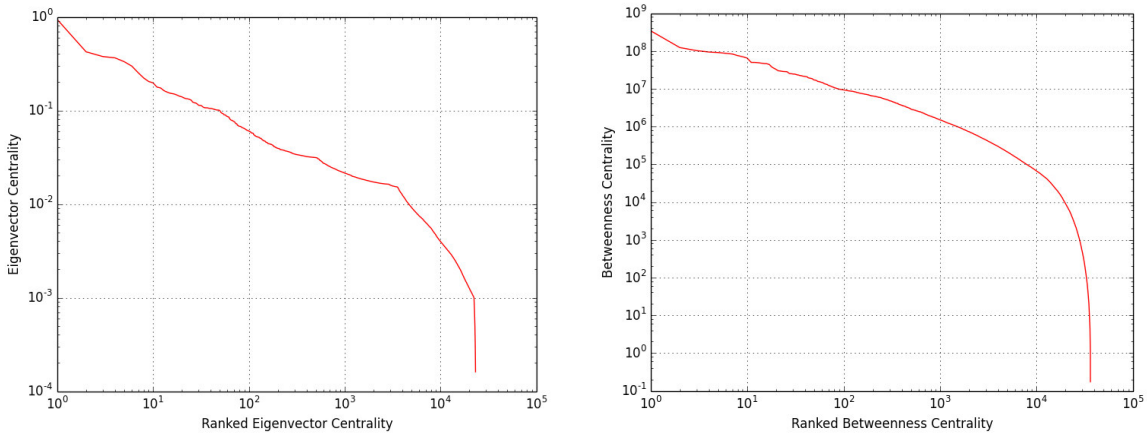
Figure 10: Hops Distributions for the FWCC and FWCCR Graphs

3.1.6 Centrality Measures

Information about the relative importance of nodes and edges in a graph can be obtained quantitatively through the centrality measures. Two of the widely used centrality measures in network analysis are the eigenvector centrality and betweenness centrality. The eigenvector centrality assigns relative scores to all nodes in the network based on the principle that connections to high - scoring nodes contribute more to the score of the node in question than equal connections to low - scoring nodes. It thus measures the importance of nodes in a network. The Google's PageRank is type of the eigenvector centrality measure. Betweenness centrality on the other hand measures the extent to which a vertex in a network lies on the pathes between other vertices. The eigenvector and betweenness centralities of the FWCC network are shown in Figure 11. Both centralities measures are right-skewed with the eigenvector centrality in Figure 11(a) having the tail of its distribution approximately follows that of a power law but the distribution rolls off for vertices with low centrality values. The same scenario can be seen in Figure 11(b) for that of the betweenness centrality. The two Figures also revealed that a good number of nodes in the FWCC network have very high eigenvector and betweenness centralities, meaning that there are very important words whose removal could disrupt the network.

4. Further Analysis and Application

In this section the relevance of the various metrics outlined in this paper are analyzed and some application areas are also given. The words distribution in the corpus is verified to follow the Zipf's law with the 3 most frequent words been words directly related to the subject matter. The 4th frequent word in the corpus is the which is known to be the most frequent word in every English text. Also, nearly 3% of the corpus vocabulary accounts for almost 90% of the words and the 3-words length account for about 22% of the text. These statistics have shown that text can be characterized and the distributions of words map out to get their pattern within a corpus. These statistics further have a lot of importance in the development of data encryption algorithms. For example, majority of words in text copra are known to be short while few words are very long as shown in Figure 6.



(a) (b)

Figure 11: Eigenvector and Betweenness centrality measures on a double logarithmic scale

Therefore knowledge of these statistics can help us develop efficient data encryption algorithms.

Secondly the lexical diversity of the corpus is seen to be very low as compare to that of the US-IAC.

This low lexical diversity is due to high re-tweeting of the same messages. This can be inferred from Figure 3, for the 3rd most frequent word in the corpus is rt which stands for re-tweet. The low corpus diversity can therefore be used as a gauge to measure the level of "grouping thinking" in social media space.

Another interesting phenomena is the burstsiness nature of text and the distribution of the word the within the FWCC. Averagely, the word "*the*" appear consecutively every 43 times in a row but majority of the words appear far above this average. This burstsiness nature of the word ultimately lead to a power-law distribution as shown in Figure 5. The most frequent word in most English text is also confirmed to have a power-law distribution. This finding also has relevance in information retrieval, data encryption and encoding.

The bigram network also shows a lot of complex network properties. For example, the degree distribution of the network is verified to follow power-law with a coefficient value of 2.14 and also confirmed that the distribution could not have been due to a random process. The joint degree distribution of the bigram network confirms that text network is different from social networks where like degrees connect to like degrees exhibiting the homophily effect but rather is like technology networks where low degree nodes are connected to high degree nodes [17].

The distance distribution among words is seen to be very short and scaled logarithmically with the number of nodes |V|, exhibiting the small world property. For the average path length is 4.78 which is within the range of other complex networks average distances such as Social, Technological and Biological networks [18] and this is a common characteristic of scale free networks. This again confirms that text can be modeled as a complex network.

The other important observation in the network metrics is the eigenvector and betweenness centralities.

The distributions of these centralities is shown to follow power-law with an exponential cutoff which is common characteristics of most eigenvector and betweenness centralities measures of complex networks.

It is therefore revealing to node that, the bigram network possess most if not all of the characteristics of complex networks. These findings could be another awakening call to analyzing the topological characteristics of text and comparing the results to that of other network metrics.

And these findings could help improve several other fields related to information science. For example, these findings could help design kernels that allow machine learning algorithms such as support vector machines to learn from string data and find likely candidates for the correct spelling of a misspelled word base on the word length or degree and also improve compression algorithms. These findings can also be applied to pattern recognition systems, speech recognition systems, optical character recognition(OCR) systems, Intelligent Character Recognition (ICR) systems, machine translation and in information retrieval systems, e.t.c.

5 Related Work

Social media content generation is vast and contains a lot of data in an unstructured, noisy and dynamic format. In the advent of social media, there have been growing interest in mining and studying the user-generated content of the social media. Among these include the work of G. Pritam and H. Liu [19]. They, in the form of a tutorial outlined the basics of social media and data mining with illustrative examples from different social media platforms. Their work highlighted the importance of social media and text mining in various domains

ranging from humanitarian assistance to disaster relief. Also G. Chakraborty *et al*. [20] in their book catalogue various text mining methods using SAS® as a tool. The text also discussed into detail the importance of text mining in today's big data era. Link-based text classification is studied by Q. Lu and L. Getoor [21]. They proposed a logistic regression framework for modeling link distribution using web and citation data sets. The model is accessed to be able to classify text with high accuracy.

CM. Tan et al. [21] also in their paper, outlined an efficient algorithm for text categorization based on bigrams. They test the efficiency of the algorithm using the McNemar test, and other measurements such as F1 measure and break-even points measure. A work that is closely related to our own is that of V. Batagelj *et al.* [22] paper "network analysis of text". The paper shows how text data can be structured as a network based on several selection criteria using ideas from graph theory and Pajek software as a visualization tool. Their work shows how text network can be extracted from text and also as a case study they show how text network is used to study the Reuters Terror news Network based on 9-11 attack on U.S. M.E.J. Newsman [6] studied power-laws, Pareto-distributions and Zip's law in many domains. He observed that the cumulative distribution of the number of times that words occur in a typical piece of English text follows a power-law. His work was based on the novel written by Herman Melville, Moby Dick.

Our work however, went beyond the sematic analysis of text to demonstrate how text can be seen as a complex network. We also used complex network tools to extract some complex network properties such as degree distribution, centralities measures, hub and authority, shortest path length distribution among others in order to prove that text data can be treated as a complex network.

6. Conclusion and Future work

The social media space is a enormous library of data. However a composite model that captures quantitative patterns and correlations in social media space is still an open question. To this end, the paper shows how the 140 character words of tweets though messy contain important patterns and inter-relationships by quantitative analysis. The FIFA World Cup Corpus (FWCC) is confirmed to have several important textual statistics. The corpus is observed to obey the Zipf's law of words distributions. The corpus further showed that 3-words length were the majority and this demonstrates users ingenuity in trying to maximize the 140 character upper bound in Twitter. Moreover, a bigram generator is proposed that modeled text data as graph and then used graph theory as a tool to analyze it. The bigram network/graph analysis revealed most complex network properties such as degree distribution that obeys power-law with degree exponent γ value of 2.14 and words distances from each other scaled logarithmically with the number of nodes in the network. Also the eigenvector and betweenness centralities values are shown to exhibit power-law distribution with an exponential cutoff toward the tails of the distributions.

These findings have a lot of applications in other fields such as data compression algorithms, speech recognition, information retrieval, data encoding, e.t.c. Also the findings show a high level of similarity between text network and that of other complex networks known in literature. In the future, we intend to investigate the network characteristics of corpora from different languages and then compare these characteristics with complex networks from different domains. This will help to know the universality of these metrics or otherwise in text network. Also we shall extend this work to look into the application of these finding in predictive policing and information security using social media space.

References

[1] A. M. Kaplan and M. Haenlein (2010). Users of the world, unite! The challenges and opportunities of social media. Business Horizons 53(1):59-68.
[2] Andreas H., Andreas N., and Gerhard P.(2005). "A Brief Survey of Text Mining." In *Ldv Forum*, vol. 20, no. 1, pp. 19-62.
[3] http://en.wikipedia.org/wiki/Data_mining#cite_note-acm-2 [Accessed on 08-04-2015]
[4] W. F. Nelson, and H. Kucera (1982). "Frequency analysis of English usage". Lexicon and grammar Boston Houghton Mifflin.
[5] JB. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and Aiden , E., L. (2011). "Quantitative analysis of culture using millions of digitized books". Science, 331:176-182, doi:10.1126/science.1199644.
[6] Newman , M., E., J. (2005). "Power laws, Pareto distributions and Zipf's law." Contemporary physics 46.5 :323-351.
[7] Olmer, P. (2008) "Knowledge Discovery". Publisher: CERN
[8] Wikipedia: *http://en.wikipedia.org/wiki/2014 FIFA World Cup* [Accessed on 16/09/2014]
[9] Russell, M., A. Mining the SocialWeb: Data Mining Facebook, Twitter, LinkedIn, Google+,GitHub, and More (2013). "O'Reilly Media, Inc.", Page 365.

[10] Zipf, G., K. (1932). “Selective Studies and the Principle of Relative Frequency in Language”.
[11] Newman, M. E. J. (2001). “The structure of scientific collaboration networks”. Proceedings of the National Academy of Sciences ; Vol. 98, No. 2, pp. 404-409.
[12] Re´ka, A., Jeong, H., and Baraba´si, A., L.(1999). “Internet: Diameter of the world-wide web”. Nature; Vol. 401, No. 6749, pp. 130-131.
[13] Cohen, R., Erez, K., ben-Avraham, D., and Havlin, S. (2000). “Resilience of the Internet to random breakdown”, Physical Review Letters 85 , 4626C4628.
[14] Callaway, D. S., Newman, M. E. J., Strogatz, S. H., and Watts , D. J. (2000). Network robustness and fragility: percolation on random graphs, Physical Review Letters 85 , 5468C5471.
[15] Van Steen, M. (2010). Graph Theory and Complex Networks. An Introduction. page 144
[16] Alstott,J., Bullmore, E., Plenz, D. (2014). powerlaw. A Python package for analysis of heavy-tailed distributions. PLoS ONE 9(1): e85777.
[17] Newman, M. E. J. (2002). Assortative mixing in networks, Phys. Rev. Lett. 89: 208701.
[18] Newman , M. (2010). Networks: an introduction. Oxford University Press, page 243.
[19] Pritam, G. and Liu, H. (2012). “Mining social media: a brief introduction.” Tutorials in Operations Research 1, (2012).
[20] Chakraborty, G., Murali, P., and Satish, G. (2013). Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS. SAS Institute.
[21] Lu, Q., and Getoor, L. (2003). “Link-based classification.” ICML. Vol. 3.
[22] Batagelj, V., Mrvar, A., and Zaveršnik, M. (2002). Network analysis of texts. University of Ljubljana, Inst. of Mathematics, Physics and Mechanics, Department of Theoretical Computer Science.

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

Prospective authors of journals can find the submission instruction on the following page: http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

MORE RESOURCES

Book publication information: http://www.iiste.org/book/

Academic conference: http://www.iiste.org/conference/upcoming-conferences-call-for-paper/

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar