Big Data Exploration in New Media: Trends and Methodological Approaches

Bisallah H. I*, Olumide O*, Aminat A*

*Department of Computer Science, University of Abuja, Nigeria hashim.bisallah@uniabuja.edu.ng

Abstract: This paper focuses on drawing trends and methodological evidences from previous studies in big data field considering the fact that big data helps individuals, governments and businesses understand the essence of collecting large data and processed it towards effective decisions making having identified many patterns within it. The review of 32 journal articles shows that most of the scholars developed own and used Internet-enabled software for exploration of big data in the new media. This indicates an existing problem in terms of availability of the right data collection and analysis tools. The review also reveals that majority of the scholars found it difficult to select appropriate samples using probability procedure. The attention was largely shifted to non-probability procedure. On the research trends, the scholars exclusively focused on relationship and politics, revealing behavioural patterns in the network, discourse focus and network connectivity, demographics as determinant of interactions and connections, patterns and messages speed at the expense of news and business areas. These issues and trends have created gaps for potential scholars in the field to conduct studies that would contribute to the knowledge base in the big data field.
[Bisallah H. I, Olumide O, Aminat A. **Big Data Exploration in New Media: Trends and Methodological Approaches.** *Academ Arena* 2019;11(6):21-32]. ISSN 1553-992X (print); ISSN 2158-771X (online). http://www.sciencepub.net/academia. 5. doi:10.7537/marsaaj110619.05.

Key words: Big Data, New Media, Methodology, Trends, Network, Demography

1.0 Introduction

New media or technologies comprise the Internet, mobile phones, email, SMS (Short Message Service), IM (Instant Messenger) among others that are helping people and businesses to reach one another using a digitalized process. Since its emergence, they have been embraced and still being used to create new things and disseminate necessary information within short period. For instance, conventional mass media no longer depend on heavy investments and large distribution costs because new technologies, especially the Internet has made distribution less expensive and more affordable (Stenberg, 1997). Social media is a form of computer-mediated communication which enables people to exchange photos and videos, share news stories, post their thoughts on blogs, and participate in online discussion. It is an offshoot of word of mouth that scaled up by leveraging the pervasiveness of the Internet (Sajithra and Patil, 2013). As web-based technologies that enhance the social architecture of a community and value of personal interactions continue to surface, social media have led to the collective creation and sharing of network data which represent the nodes and edges (Smith, Shneiderman, Milic-Frayling, Rodrigues, Barash, Dunne, Capone, Perer and Gleave, 2009; Guy, 2012; Sajithra and Patil, 2013). The collective creation and data distribution have caught attention of many scholars in recent times (Meraz, 2011; Lim and Datta, 2012).

Figure 1: History of Social Media
Source: Sajithra and Patil (2013)

On these new media, people and organisations are creating and receiving various data every day in the course of conversing with one another. This has resulted to big data, which has been trending in literature and books since 2012. The central focus of big data is to let individuals, governments and businesses understand the essence of collecting large data and processed it towards effective decisions making having identified many patterns within it. Collection and processing of large data are done with modern and sophisticated computing techniques than using the traditional ones. In other words, the two stages require many areas of technology. Huge volume, high velocity and extensible variety of data have been seen as the specific features of big data. Big data could be structured one that is relational data, semi-structured which entails XML or unstructured one which has word, PDF, text, media logs as its variety.

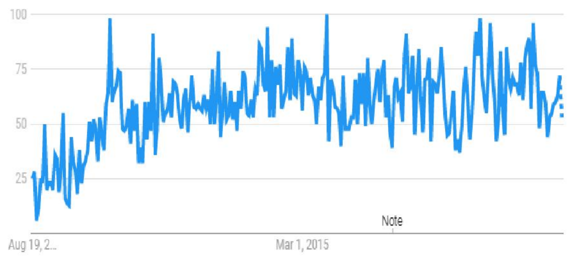

Figure 2: Big Data Trends in Books and Literature 2012-2017
Source: Google Trends, 2017

1.1 Statement of the Problem

Evidently, big data has reached a point which majority of businesses especially the Fortune 1000 firms cannot do without it. Businesses are now viewing big data as very important or critical to their operational effectiveness and attainment of corporate goals (New Vantage Partners, 2016). Exploration and analysing a large dataset is becoming increasingly common with the main aim of offering different patterns of perceiving and manipulating information as well as knowledge extraction and inference (Bikakis and Sellis, 2016). In specific terms, businesses and governments are now finding, visualizing and understanding big data towards improved decision making. Based on this conceptual background, this study intends to assess existing empirical studies with the specific reference to big data exploration in the new media. This becomes imperative since big data keeps evolving in business and non-business environment. Beyond this, methodological issues and gap in knowledge that need to be filled by new researchers would be revealed.

1.2 Purpose/Aims of the Study

The study specifically investigates the following;

a. The interests or areas of specialization of the previous researchers or scholars in the big data exploration within new media or technologies.

b. Research approaches adopted by the scholars towards understanding the kind of problems they studied or investigated.

c. Theories that helped the scholars in setting their findings into new perspectives, having adopted and tested specific propositions and assumptions.

d. Emerging dominant findings from the previous studies that revealed big data at micro and macro levels of research participants.

1.3 Significance of the Study

The outcomes of the study will be useful and helping future researchers in furthering the big data knowledge base especially its exploration in various new technologies. Future academic researchers will gain research trends and methodological insights of the existing studies. These insights will give specific areas that need further studies and averting replication of the existing ones.

1.4 Delimitation of the Study

The review is delimited to peer-reviewed academic computer science related journals published from 2011 to 2017. Conference proceedings, technical reports, book chapters, periodicals were not included for the review.

1.5 Research Questions

The following research questions guide the review process;

1. What are the focus areas of big data exploration in the new media from 2011-2017?
2. What are the dominant methods for big data exploration in the new media during the period?
3. What are the theoretical trends within the period?
4. What are the main findings of the studies during the period?

2.0 Literature Review

2.1 New Media

As captured under the introduction, new media give people and organisations opportunity of sharing or exchanging information on a specific discourse or topic with the main aim of creating and building social capital. Beyond the social relationship building and maintenance, businesses and governments are not left out in the usage of various new technologies such as the Internet, Facebook, Twitter, Instagram, YouTube among others. Out of these technologies, social media usage are gaining more trend than others. This has been premised on the fact as espoused by Guy (2012) while referring to Bryer and Zavatarro (2001) that "Social media are technologies that facilitate social interaction, make possible collaboration, and enable deliberation across stakeholders" (p. 327).

2.2 Big Data and Exploration

The use of different new media for business and non-business purposes has been generating a lot of data set and processed into insights towards effective strategic decision making. Big data, as a concept, has been defined by scholars and researchers using different approaches and perspectives. Bhosale and Gadekar (2014) see big data as a large data set which requires new architecture, techniques, algorithms and analytics before it could be managed, extracted and mined hidden knowledge from it. Bhosale and Gadekar (2014) add that Big data as a term refers to data sets or combinations of data sets whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult to be captured, managed, processed or analyzed by conventional technologies and tools, such as relational databases and desktop statistics or visualization packages, within the time necessary to make them useful. While the size used to determine whether a particular data set is considered big data is not firmly defined and continues to change over time, most analysts and practitioners currently refer to data sets from 30-50 terabytes (10 12 or 1000 gigabytes per terabyte) to multiple petabytes (1015 or 1000 terabytes per petabyte) as big data. To capture insights within the big data in terms of its velocity, variety, veracity and volume, application, computing and infrastructure layers are needed. This should be applied from the infrastructure perspectives before reaching computing to application (Bhosale and Gadekar, 2014).

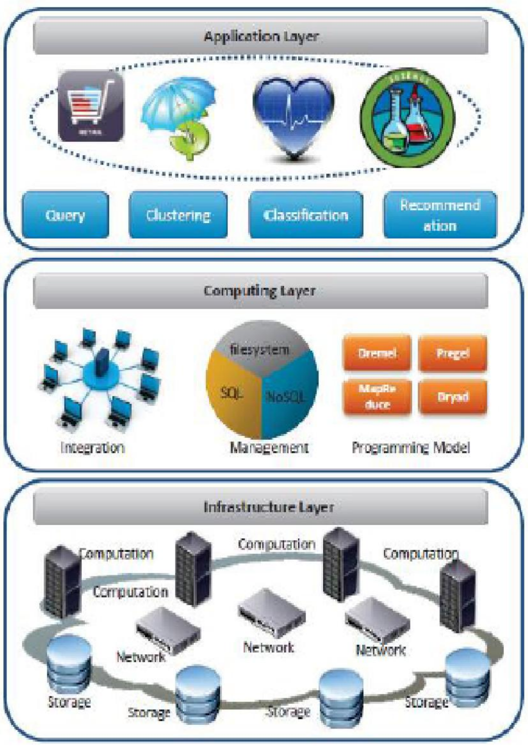

Figure 3: Layered Architecture of Big Data System
Source: Bhosale and Gadekar (2014)

In the literature, big data exploration has been ascribed with data mining, knowledge extraction, information discovery, information harvesting, data archaeology and data pattern processing. However, data mining remains the mostly used by scholars especially statisticians, data analysts, and the management information systems (MIS) communities (Fayyad, Piatetsky-Shapiro, and Smyth, 1996). NoSQL (largely being translated as "not only SQL"), Hadoop among others have been described as leading software for capturing, curating, visualizing and analysing the fast-paced volume expansion of the unorganized data (Sharma, undated).

3.0 Methodology

This chapter detailed specific procedures employed for carrying out the study under the headings; research design, population of the study, data collection and analysis procedures.

3.1 Research Design

Quantitative research approach was employed to execute the study. Content analysis and Systematic review were specifically adopted. As a quantitative research method, content analysis helped the researcher in analysing manifest content of previous studies on big data research exploration using content categories as instrument in line with the research questions that guided the study. Systematic review is adopted as parts of the research designs because it assisted in mapping out existing findings in order to avoid duplication of areas previously researched on and identify gaps in knowledge that need additional research from scholars (Kofod-Petersen, 2014).

3.2 Population of the Study

Materials in form of published journals constituted population of the study. These are the peer-review journals that entailed big data and exploration articles written by scholars in computer science and computer-mediated communication fields.

3.3 Data Collection and Analysis Procedures

To gather necessary data through the instrument, Google Scholar, OpenDOAR, Directory of Open Access Journal and RefSeek were the search engines that helped in retrieving relevant journals for analysis. These search engines were selected because they provide free access journal articles which assisted researcher in reducing huge financial requirements associated with the paid academic search engines. Big data and big data exploration were the two main keywords developed for retrieving the journals from the chosen search engines. These keywords were expanded into phrases to help in gathering more journals. Big data and new media, big data and social media, big data and the Internet, big data and news websites were the five phrases crafted from the 'Big Data' while Big data exploration findings, research design for big data exploration, big data exploration in the new media, big data exploration in social media, big data exploration in the Internet, big data exploration in the news websites emerged from 'Big Data Exploration'. Suffice to note that the first twenty hits emanated from each keyword and phrase was pinpointed for journals retrieval. The gathered data were subjected to simple frequency percentage count of descriptive analysis and synthesis using comparative constant method.

Categories Operationalization

a. Focus: This is the summarised purpose or objective of each study reviewed under the themes; business, politics, economy, relationship, news, others.

b. Method: This indicates research approach or design employed by the researchers (s) in executing their studies' focus. These include exploratory, survey, computational, content analysis among others.

c. Theories: These are the framework works in terms of propositions that helped the researcher (s) in verifying their findings. Network, graph and other likely emerged theories were considered.

d. Sampling Procedure: This represents specific sampling technique (s) used by the researcher (s) for the selection of samples from the population they studied. The technique could be chosen from the probability or non-probability approaches.

e. Data Collection and Analysis: These are the tools or techniques adopted by the researcher (s) for gathering of relevant data and analysed accordingly. NodeXL, NoSQL, Hadoop, MangoDB among others were measured.

Table 1: Search Terms/Phrases

Keywords	Extended Term/Phrase
Big Data	Big data and new media, big data and social media, big data and the Internet, big data and news websites
Big Data Exploration	Big data exploration findings, research design for big data exploration, big data exploration in the new media, big data exploration in social media, big data exploration in the Internet, big data exploration in the news websites

Table 2: Inclusion, Exclusion and Quality Criteria

Criteria Identification	Inclusion Criteria	Exclusion Criteria
RQ1	Title, Abstract, Objectives/Purpose Section	Non-availability of the full text online; technical reports, book chapters, periodicals, conference papers
RQ2	Methodology Section	
RQ3	Abstract, Introduction and Theoretical Section	
RQ4	Abstract, Result Section	

Table 3: List of Search Engines

S/N	Search Engine	URL
1.	Google Scholar	https://scholar.google.com/
2.	Directory of Open Access Journal	https://doaj.org/
3.	OpenDOAR	http://www.opendoar.org/search.php
4.	RefSeek	https://www.refseek.com/

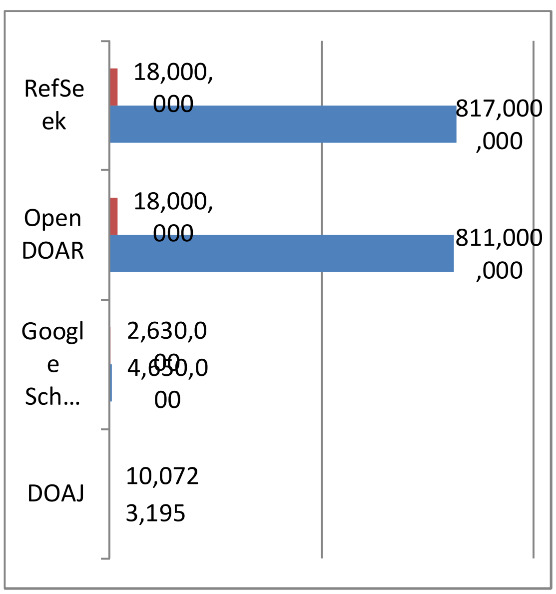

Figure 4: Keywords

4.0 Results And Discussion

This chapter focuses on the presentation, analysis and discussion of the gathered data. It encompasses the description of the reviewed journals, specific findings for research questions that guided the review and synthesis of the dominant findings of the journals.

4.1 Description of the Journal Articles Reviewed

Out of the two keywords used in searching relevant publications, big data generated more hits than big data exploration. RefSeek gave 817 million hits, which is the highest followed by OpenDOAR search engine. Under the big data exploration, OpenDOAR and RefSeek also led Google Scholar with 18 million hits accordingly. For the extended terms and phrases, OpenDOAR and RefSeek were also found to give more publications relevant to big data exploration in the New Media. When big data is combined with New Media, RefSeek gave 763 million hits while OpenDOAR disclosed 620 million hits. Big data and social media generated over 300 million results for both OpenDOAR and RefSeek. The use of big data and the Internet revealed 219 million hits for OpenDOAR and RefSeek respectively. Big Data and News Website as a phrase led to the generation of 747 million hits for RefSeek and 744 million for OpenDOAR. From these results, a total of 131 articles published in journals were retrieved using the formulated inclusion and exclusion criteria. Out of these 32 articles met the final criteria.

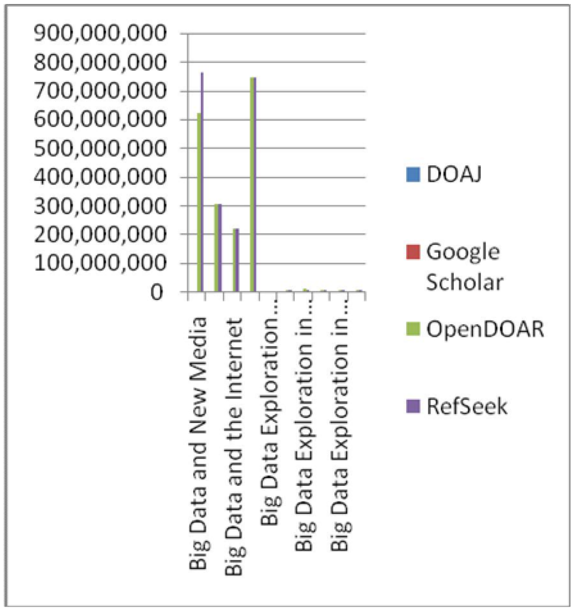

Figure 5: Extended Term/Phrase

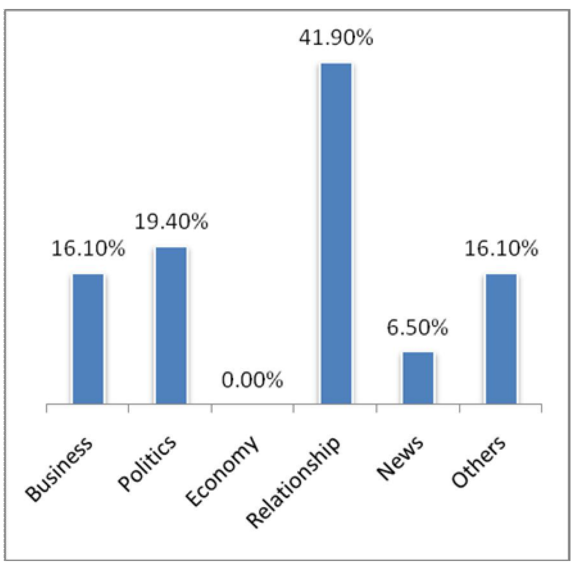

Figure 6: Research Focus

4.2 Analysed Data and Results Presentation

4.2.1 Focus areas of big data exploration in the new media from 2012-2017

Relationship and politics were the main focus areas of most scholars of the articles. Over 41% of the scholars carried out their studies with the intention of finding solutions to issues among people in relation with the New Media. More than 19% of the articles were written around political issues or problems. Business and others purposes were tied on 16.10% while News had the least (6.50%).

4.2.2 Dominant methods for big data exploration in the new media during the period

Having understood the main areas of the scholars, attempt was made to know specific research methods adopted. Computational (35.10%), Content Analysis (21.60%) and Exploratory (18.90%) designs were mostly used by the scholars. When sampling techniques of the scholars were reviewed, the results indicates adoption of non-probability sampling procedure (79.30%) followed by probability (20.70%). Against the researcher's expectation, Hadoop as one of the useful tools for big data collection and analysis was used less (6.90%) than others (82.80%). Others category encompassed the use of personalized (developed or created by scholars) and established (using existing free open) tools. For instance, Application Programming Interface was used singlehanded or combined with other tools such as MongoDB, Webometric and Snowcrawl (Thelwall and Sud, 2011; Bruns, Highfield and Burgess, 2013; Crampton, Graham, Poorthuis, Shelton, Stephens, Wilson and Zook, 2013; Bail, 2014; Poorthuis, Graham and Zook, 2014; Jang and Hart, 2015; Yu and Wang, 2015; Willis, Fisher and Lvov, 2015; Liu, Preoţiuc-Pietro, Samani, Moghaddam and Ungar, 2016; Zhou and Xu, 2017).

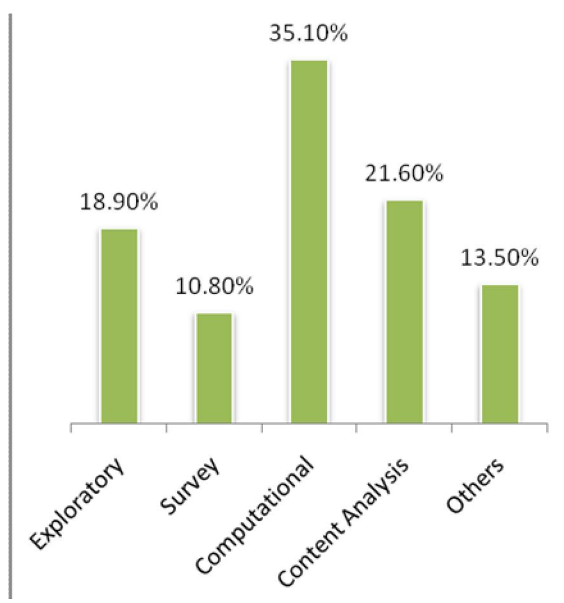

Figure 7: Research Design Used

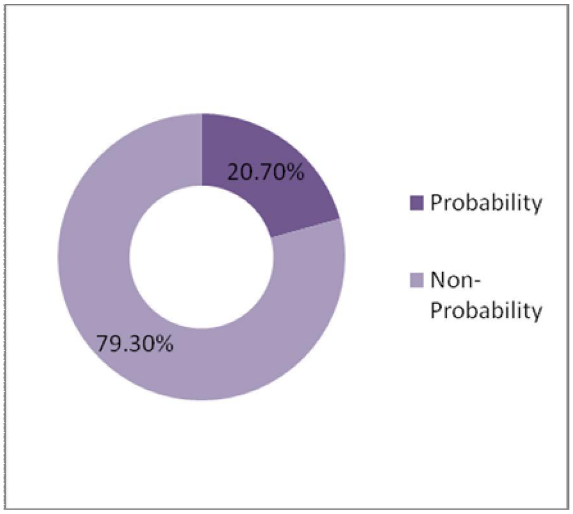

Figure 8: Sampling Techniques Used

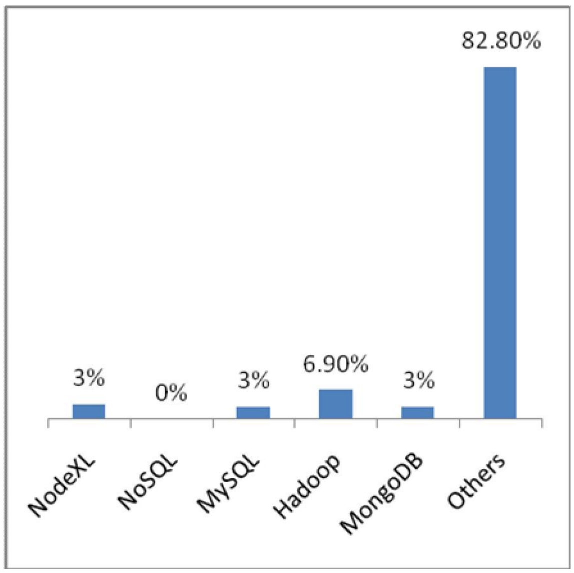

Figure 9: Data Collection and Analysis Used

4.2.3 Theoretical trends within the period

Majority of the scholars (others=60.00%) adopted and adapted propositions of theories and models such as conspiracy, selective exposure, psychological, semiotic, agenda setting, evolutionary, communication privacy management, disposition, framing. Big Five Personality, latent space and Territory-Place-Scale-Network (TPSN) models were used by some scholars. Apart from these theories and models, Network (30.00%) and Graph (10.00%) were also used.

4.2.4 Main findings of the studies during the period

The reviewed studies established a number of significant findings premised on topical issues discussed or reported using posts, tweets, comments, replies, retweets among other forms of communication in a digitalised environment which resulted to social relationship and connection with peers having similar or different characteristics and philosophies.

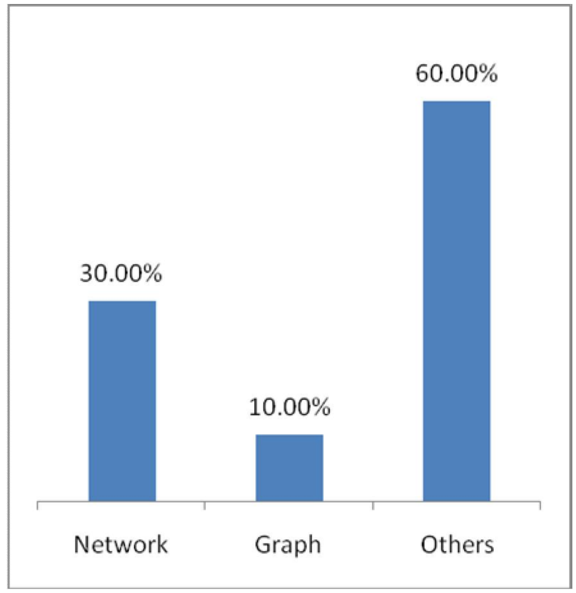

Figure 10: Theoretical Underpinnings Used

4.2.4.1 Behavioural Patterns in the Network

Findings indicated that during watching of live football match fear and anger as the most common negative emotions increased when the opponent team scored and decreased when own team scored (Yu and Wang, 2015). It has also been found that joy and anticipation manifest mostly in fans' tweets than negative emotions in line with the goal results and real time circumstances (Yu and Wang, 2015). Kramer, Guillory and Hancock (2014) have earlier found members of a social network's emotions as predictor of others expression of negative or positive emotions, indicating others' positive experiences constitutes a positive experience for people. As negative messages grow in a network, members of the network are eager to check updates regularly with the intent of maintaining established social capital. Visibility, connectivity and persistence as privacy violations determinants afforded individual member opportunity to have constant social comparison to other members and triggered jealousy, anxiety, and other negative emotions (Fox and Moreland, 2015). However, as soon as positive expressions dipped in a network, members generated fewer posts and more negative posts whereas the reduction of negative messages led to more positive posts (Kramer, et.al.2014). In a political network, information is exchanged basically among the members with similar ideological orientations or preferences rather than a network of current events (Barberá, Jost, Nagler, Tucker and Bonneau, 2015). Findings have showed that members self-defined their identity rather than identity imposition by outsiders whether negative or positive expressions are being traded (Lewis, Zamith and Hermida, 2013). At the business level, lower market share was found to predict company's high degree of using and monitoring social media activities especially responding to customers' comments quickly (He, Zha and Li, 2013).

4.2.4.2 Discourse Focus and Network Connectivity

Issues and happenings within socio-economic and political domain are mostly exchanged among the users of different new media and examined by the scholars. For instance, Caton, Hall and Weinhardt (2015) found that Facebook is used mainly as a medium for promoting individual (political) agendas. This aligned with the Allcott and Gentzkow's finding (2017) that fake news was both widely shared and heavily tilted in favour of Donald Trump during United States' Presidential Elections. Within political virtual sphere, specific controversial topics had been found to attracting conservative and liberal clusters whereas broader topics reflected conservative sentiment (Himelboim, McCreery, and Smith, 2013). Negative event affecting a large number of people with associated messages towards helping the victims and potential ones are highly distributed in social media while messages related to insignificant health issues are unspecific and topically diverse in real-life settings (Yang, Kiang and Shang, 2015). From the social perspective, negative happenings such as illegal killing or shooting, preference for music or lifestyle discussion on social media could begin as a national conversation before transforming into a polarized exchanged and creating tense digital environment. Such environment could be set up by religion differences. This has been linked with the public nature of conflict on social media mostly Facebook (Thelwall and Sud, 2011; Fox and Moreland, 2015; Barberá, et.al. 2015).

As members of networks engaged each other on the topics or issues being discussed various connection patterns occur. In their study, Himelboim, et.al (2013) found a strong relationship between a message's political orientation and the orientation of the content it linked to (Pearson Chi-Square = 1011.96, $p < .001$). For instance, the scholars discovered that 81.7% of messages with liberal political orientation linked to content with liberal political orientation, 75.4% of conservative messages linked to conservative content, and 76.1% of messages lacking an obvious political orientation linked to content without clear political orientation. However, difference was found in the linkage of content with conservative and liberal orientation in the emerging clusters in the network (Himelboim, et.al.2013). When interaction level is measured it has been discovered that there are factors that accounted for variation among the members. In a political network examined by Caton, et.al (2015) interactions among politicians are relatively low: 3,883 occurrences (0.23%) across all profiles while politicians' interaction with their followers captured 385,936 bi-directional edges when text-based and likes interactions were considered. In a related study, which was conducted on the users' level alone, it was found that 325,771 of the 960,499 twitter users during two sporting activities appear more than once which signifies interaction with the content produced by the approximately 1,000 accounts (Willis, Fisher and Lvov, 2015). In a hospitality business context, link between guest experience and satisfaction are inherently connected with the Hybrid and Deals as the dominant factors (Xiang, Schwartz, Gerdes and Uysal, 2015). The connection (citizens and group) could, however, be threatened when big data are embedded in personalized marketing and content production via information, argumentation, empathy, and celebration as members of a shared social and civic space (Couldry and Turow, 2014). In the relational connections, existing scholars have also discovered that social networks enabled people to develop ideas and propagate them with the intent of transferring knowledge to others disproportionately when making negative comments while positive comments usually attracted few replies or reactions (Thelwall and Sud, 2011; Crampton, et.al.2013). Beyond relational connection, Willis, et.al (2015) also contributed to the ongoing discourse by revealing similarity of retweeting rates of @BBCSport and @london2012 accounts while the accounts played different roles within the network. It was discovered that @BBCSport has relatively high Betweenness centrality which made it a provider of information to specific audience or demographic within the entire network more than @london2012. Park and Leydesdorff (2013) examined existed network among the scholars in the United States and the rest of the world in terms of co-authorship in big data research. Findings showed that the US had the largest number of papers co-authored with other countries, and its degree centrality was 4.450, accounting for 17.5% of the network, followed by the Germany, United Kingdom., France, Australia, the Netherlands, China, Denmark, and Canada. Taiwan and South Korea also joined China as countries that collaborated with the US' scholars the most. While the US made some strong ties with Asian countries (e.g., Singapore, South Korea, China, and Taiwan), it has a relatively sparse network with European countries (e.g., Switzerland, Denmark, Spain, The Netherlands, France). Many countries are weakly connected with Norway, Denmark, and Sweden with a UK's group formation (Park and Leydesdorff, 2013).

4.2.4.3 Demographics as Determinant of Interactions and Connections

From personalized to geographical characteristics, users of social and new media interacted and connected on issues or events interested them. This was verified empirically by some of the studied reviewed. At personal level, educational status, age and total media consumption have been found as predictors of the extent to which people believe (true or false) news headlines (Allcott and Gentzkow, 2017). When it is about exchanging business ideas females have been found to post topics related to small business on twitter than males (Balan and Rege, 2017). On the YouTube, mildly positive comment was posted by a 29 year old male which contained 58 characters (Thelwall and Sud, 2011). Align with Allcott and Gentzkow's finding (2017), Eastin, Brinson, Doorey and Wilcox (2016) earlier study has showed that privacy concerns of collection, awareness and location were not significant predictors of mobile commerce activity whereas control and unauthorized access significantly predicted the activity. Overall, the regression explained a significant amount of variance (F (11,404) ¼ 29.24, p <.01, R ¼ 0.67, Adjusted R2 ¼ 0.43) (Eastin, et.al.2016). In their study, Liu, Singh and Srinivasan (2015) discovered Google Trends, Wikipedia views, IMDB reviews, and Huffington Post news as the weak predictors of TV show demand because users tweet about TV shows before, during, and after a TV show, whereas Google searches, Wikipedia views, IMDB reviews, and news posts typically lag behind the show.

Findings have also been generated based on geographical locations of the users of the new media. In a study conducted by Zhou and Xu (2017) with the intent of knowing geographical bases of the people who tweeted about the Pope's visit to the United States, it was discovered that tweets were enormous in line with the places Pope visited in the country. Findings revealed that most of the tweets emanated from DC where the Pope first visited. As soon as the Pope moved to the New York on another day the number of tweets increased significantly from the people living in the city as well as in Philadelphia when Pope moved there. The study specifically established that the movement of the tweets largely conformed to the itinerary of the Pope (Zhou and Xu, 2017). In similar study, Shelton, Poorthuis, Graham and Zook (2014) found a significant concentration of Sandy-related tweets along the eastern seaboard of the US, especially in those places that were most affected by the storm, with approximately 30% of all Sandy-related tweets being located in the New York City metropolitan area. In the Northeastern part, clustering of tweets in the region was the material manifestation of Hurricane Sandy. In the region, the Lower East Side and Coney Island had significant levels of tweeting activity. However, some of the hardest hit places also had relatively little tweeting activity. The scholars found that physical distance has no significant relationship with the relative level of tweeting activity about Hurricane Sandy (Shelton, et al., 2014). Earlier study has showed that tweets could be concentrated sparingly when TV social response in a real-time manner is the focus (Hu, Wen, Gao, Chua and Li, 2013).

Beyond geographical concentration of tweets, some of the scholars have also examined twitter users' race or ethnicity within the network they belonged. In the examination of #egypt and #libya during the Africa-Arab nations' revolution Bruns, Highfield and Burgess (2013) discovered a larger group of Arabic-speaking users as participants in the #egypt discussion than in #libya. There was also a substantial shift over time of the tweets. Apart from the use of Arabic language for discussion purposes, a significant number of users were found using Latin characters and established extent to which information exchanges are able to bridge existing language divides (Bruns et.al. 2013).

Considering framing a global issue across countries, Jang and Hart (2015) discovered that tweets generally mirror much of the controversy observed in the traditional media, reporting climate change. From the United States to Australia, the climate change was framed differently. The real frames were created by users from the US than the UK, Canada and Australia. Hoax frames were more prevalent in the US than in the UK, Canada, and Australia. American users were also found displaying different patterns when it came to the cause or consequence of climate change and how it should be treated. The US registered a lower ratio of cause frames than the UK, Canada, and Australia (Jang and Hart, 2015).

As information is exchanged within shortest time without locations and language creating barriers, lifestyle of the members of different networks determine patterns of reaction to the networks' focus. For instance, profile picture choice and personality traits have been established to play a significant role in social media users' emotions. Users with the agreeable and conscientious personality traits display more positive emotions in their profile pictures, while users high in openness prefer more aesthetic photos (Liu, Preotiuc-Pietro, Samani, Moghaddam and Ungar, 2016). However, the users might be inconsistent with the preferred profile picture choice and the kind of emotions they displayed. This observation could be situated within Fox and Moreland's study (2015) in which majority of Facebook users claimed that the medium was inconsequential and later recounted significant stressful or hurtful events associated with it.

Political differences or orientations have also been investigated by most of the scholars, especially in developed countries such as the US. Democrats and Republicans' reactions to issues have been the focal point of the studies (Colleoni, Rozza and Arvidsson, 2014; Allcott and Gentzkow, 2017). In Colleoni et.al, (2014) Democrats exhibit higher levels of political homophily while Republicans who follow official Republican accounts exhibit higher levels of homophily than Democrats. On a significant note, the scholars discovered high levels of homophily in the network of reciprocated followers than in the non-reciprocated network. Joining Colleoni et.al, (2014), Barberá, et.al's study (2015) showed that liberals were more likely than conservatives to engage in cross-ideological dissemination when political and non-political issues are being discussed. On reactions to political news, Democrats and Republicans are both likely to believe ideologically aligned headlines especially among the people with ideologically segregated social networks (Allcott and Gentzkow, 2017). Outside the US, social media have been established to be more responsive to the socio-economic and political issues based on the extent to which people use them to abstracting immigration, the Arab Spring and conflicts in Iraq and Afghanistan (Neuman, Guggenheim, Jang and Bae, 2014).

4.2.3.4 Patterns and Messages speed

The reviewed studied have also pointed out the patterns or manners of interactions in social network environment occasioned by different new media, frequency of the interactions and the time lag. Liu, et.al (2015) informed future researchers through their method devised for the collection of big data in the new media on the need to classify tweets on what is in them (tweets) rather than preconceived topics that may not be relevant. Crampton, et.al (2013) discovered spatial patterns that showed how tweets differ and evolved over time (aggregated in 5-minute bins). The extraction of TV social responses during the broadcast period has also revealed that audience accustomed to watch TV during the airing of final episode of favourite movies led to high posting of opinionated tweets simultaneously (Hu, et.al, 2013). Interactions between politicians and voters have also been established to record different frequency. On average, the two interlocutors have been found to interacted 2.70 times via comments, with a maximum of 1,503; 4.30 and 998 respectively for likes, and 4.45 and 1554 considering both (Caton, et.al.2015). Different from the politicians and individual members' interactions, finding has also noted no typical density of discussion on YouTube videos in the sense of the proportion of replies to other comments. This pattern resonated in the videos with few replies and with many replies (Thelwall and Sud, 2011). The speed at which messages move among the interlocutors occurs and differs significantly which is hinged on the nature of the events or topics has also been discovered (He, et.al 2013; Bruns, et. al 2013). He and others specifically found the difference to be associated with the special offers and discount given to the consumers by three Pizza's stores while Bruns and his colleagues' difference was largely on the shift from outright revolution in the Arab nations in Africa to a more long-term reshaping of the political system. Comparing corporate, pundits and athletes' accounts, Willis, et. al (2015) reported that athletes' tweets seem to remain relevant longer than those of the pundits and the corporate accounts. Report differently, Yasseri, Hale, and Margetts (2014) the definition and measurement of an average outreach factor for petitions revealed that after 24 hours, a petition's fate is virtually set in network. For instance, over 99 percent fail to get the 10,000 signatures required for an official response and only 0.1 percent attains the 100,000 required for a parliamentary debate (Yasseri, et. al. 2014). Expanding the discourse, Zhou and Xu (2017) have increased the existing knowledge on time lag of tweets or comments in a network where natural event is being discussed. The scholars' analysis of tweets during rainy day showed 8am as the peak period of tweets in the first day and second peak period occurred around 12 pm to 1pm, afterwards the amount of rain-related tweets dipped. Most of the tweets were posted during times when more outdoor transportation was needed (morning peak transportation hours, noon, and afternoon peak hours). However, the scholars did not find a significant match with precipitation levels based on the number of tweets emanated from the three major cities; Philadelphia, Washington DC, and New York City examined (Zhou and Xu, 2017). Findings from the second study of the scholars (Zhou and Xu, 2017) differed from the first study. The monitor of Pope's visit, a social event, showed 4pm, when the Pope arrived in DC as the first peak period while the second occurred at 10am when the Pope gave a speech at the Senate and House of Representatives. The researchers concluded that discussion on Twitter about this event was more intensive in the morning but less intensive in the early afternoon. Adding to the conversation, Caton, et.al (2015) have made future researchers understand that timing of posts and comments as well as daily patterns is an indication that politicians through their social media accounts tend to post on working days, whereas constituent volume shows no significant difference between weekdays and weekends. When people framed issues of global importance they were likely to generate varied levels of tweets based on the adopted words (Jang and Hart, 2015). For instance, "climate change" was found to be more pronounced than "global warming" apropos hoax frames. The ratio values for hoax frames were 1.65 (US), 1.93 (UK), 1.43 (Canada), and 1.18(Australia), revealing that all four countries were more likely to use the term ‘‘global warming’’ instead of ‘‘climate change’’ when tweets were related to hoax frames (Jang and Hart, 2015).

5.0 Summary, Conclusion And Recommendations

5.1 Summary

The review of the existing studies have established that there are methodological issues and gap need to be filled by the future researchers. Most of the scholars developed own and used Internet-enabled software for exploration of big data in the new media. The review also showed that majority of the scholars found it difficult to select appropriate samples using probability procedure. The focus was largely shifted to non-probability procedure. On the research trends, the scholars exclusively focused on relationship and politics at the expense of news and business.

5.2 Conclusion

The researchers’ developed of personal data collection and analysis tools indicate an existing problem in terms of availability of the right tools. The inability to use probability sampling procedure could be hinged on the fact that methodologists have not disentangle the best approach for the adoption probability techniques such simple random and systematic in a digitalized or computer-mediated communication research setting. The scholars’ exclusive focus on relationship and politics as the main areas of investigation at the expense of news and business could be associated with the view that users appropriated new media mostly for social capital building and maintenance, and political interactions.

5.3 Recommendations

These issues and trends have created gaps for potential scholars in the field to conduct studies that would contribute to the knowledge base in big data. The following areas are proffered;

1. Exploration of big data in news websites with specific reference to socio-economic and political issues or events.
2. The interplay between data on websites and social networking sites of news media’s websites.
3. Categorization of news patterns that dominate the websites and social media sites of the news media.
4. Readers’ connectivity in terms of geographical closeness.
5. The time lag of posts, tweets, replies and comments to news reports. This could also include determination of passive or active commenters to the stories posted on the social media sites.

References

1. Allcott, H., and Gentzkow, M., (2017) “Social Media and Fake News in the 2016 Election” *Journal of Economic Perspectives* Vol. 31, No. 2 Pp211–236.
2. Bail, C.A., (2014) “The cultural environment: measuring culture with big data” *Springer Science Business Media Dordrecht* Pp1-19.
3. Balan, S., and Rege, J., (2017) “Mining for Social Media: Usage Patterns of Small Businesses” *Business Systems Research* Vol. 8 No. 1 Pp43-50.
4. Barberá, P., Jost, T.T., Nagler, J., Tucker, J.A., and Bonneau, R., (2015) “Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?” *Psychological Science* Pp1–12.
5. Bhosale, S.H., and Gadekar, D.P., (2014) “A Review Paper on Big Data and Hadoop” *International Journal of Scientific and Research Publications*, Volume 4, Issue 10, Pp1- 7.
6. Bikakis, N., and Sellis, T., (2016) “Exploration and Visualization in the Web of Big Linked Data: A Survey of the State of the Art” Workshop Proceedings of the EDBT/ICDT 2016 Joint Conference (March 15, 2016, Bordeaux, France) on CEUR-WS.org (ISSN 1613- 0073)
7. Bruns, A., Highfield, T., and Burgess, J.E., (2013) “The Arab Spring and social media audiences: English and Arabic Twitter users and their networks” *American Behavioral Scientist*, Vol.57 No. 7 Pp871-898.
8. Caton, S., Hall, M., and Weinhardt, C., (2015) “How do politicians use Facebook? An applied Social Observatory” Big *Data & Society* July–December Pp1–18.
9. Colleoni, E., Rozza, A., and Arvidsson, A., (2014) “Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data” *Journal of Communication* Vol. 64 Pp 317–332.
10. Couldry, N., and Turow, J., (2014) “Advertising, Big Data, and the Clearance of the Public Realm: Marketers’ New Approaches to the Content Subsidy” *International Journal of Communication* Vol. 8 Pp 1710–1726.
11. Cramptona, J.W., Grahamb, M., Poorthuisa, A., Sheltonc, T., Stephensd, M., Wilsona, M.W., and Zoo, M., (2013) Beyond the geotag: situating ‘big data’ and leveraging the potential of the geoweb Cartography and Geographic Information Science, Vol. 40, No. 2, Pp130–139.
12. Eastin, M.S., Brinson, N.H., Doorey, A., and Wilcox, G., (2016) “Living in a big data world: Predicting mobile commerce activity through privacy concerns” *Computers in Human Behavior* Vol. 58, Pp214-220.

13. Fox, J., and Moreland, J.J., (2015) "The dark side of social networking sites: An exploration of the relational and psychological stressors associated with Facebook use and affordances" *Computers in Human Behavior* Vol.45 Pp 168–176.
14. Guy, R., (2012) "The Use of Social Media for Academic Practice: A Review of Literature," *Kentucky Journal of Higher Education Policy and Practice* Vol.1, Issue 2, Accessed on http://uknowledge.uky.edu/kjhepp/vol1/iss2/7.
15. He, W., Zhab, S., and Li, L., (2013) "Social media competitive analysis and text mining: A case Study in the pizza industry" *International Journal of Information Management* Vol.33, Pp464– 472.
16. Himelboim, I., McCreery, S., and Smith, M., (2013) "Birds of a Feather Tweet Together: Integrating Network and Content Analyses to Examine Cross-Ideology Exposure on Twitter" *Journal of Computer-Mediated Communication* Vol. 18, Issue 2, Pp40–60.
17. Hu, H., Wen, Y., Gao, Y., Chua, T.-S., and Li, X. (2015) "Toward an SDN-enabled big data platform for social TV analytics" *IEEE Network* Vol. 29 No. 5 Pp43-49.
18. Jang, S.M., and Ha, P.S., (2015) "Polarized frames on ''climate change'' and ''global warming'' across countries and states: Evidence from Twitter big data" *Global Environmental Change* Vol.32, Pp11–17.
19. Kitchin, R., and McArdle, G., (2016) "What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets" *Big Data & Society* January–June Pp1–10.
20. Kramera, A.D., Guilloryb, J.E., and Hancock, J.T., (2014) "Experimental evidence of massive-scale emotional contagion through social networks" *PNAS*Vol. 111, No. 29 Pp 8788– 8790.
21. Kofod-Petersen, A., (2014) How to do a Structured Literature Review in computer science Accessed on https://pdfs.semanticscholar.org/f9e7/b1f645dded dfbf702558f554dd316a7692ae.pdf
22. Lewis, S.C., Zamith, R., and Hermida, A., (2013) "Content Analysis in an Era of Big Data: A.
23. Hybrid Approach to Computational and Manual Methods, *Journal of Broadcasting & Electronic Media,* Vol. 57, No 1, Pp34-52.
24. Lim, K.H., and Datta, A., (2012) *Tweets Beget Propinquity: Detecting Highly Interactive Communities on Twitter using Tweeting Links.* School of Computer Science and Software Engineering, The University of Western Australia Crawley, WA 6009, Australia.
25. Lin, J., Gholami, M.J., (2014) "Infrastructure for Supporting Exploration and Discovery in Web Archives" *Companion*, April 7–11, Pp851-855.
26. Liu, X., Singh, P. V., and Srinivasan, K., (2015) "A Structured Analysis of Unstructured Big Data by Leveraging Cloud Computing" Accessed on https://doi.org/10.1287/mksc.2015.0972
27. Liu, L., Preotiuc-Pietro, D., Samani, Z.R., Moghaddam, E., and Ungar, L., (2016) "Analyzing Personality through Social Media Profile Picture Choice" Association *for the Advancement of Artificial Intelligence.*
28. Meraz, S., (2011) "The fight for 'how to think': Traditional media, social networks, and issue interpretation" *Journalism* 12(1):107-127 DOI: 10.1177/1464884910385193.
29. Neuman, W.R., Guggenheim, L., Jang, S.M., and Young, S.B., (2014) "The Dynamics of Public.
30. Attention: Agenda-Setting Theory Meets Big Data" *Journal of Communication* Vol. 64 Pp193–214.
31. New Vantage Partners (2016) *Big Data Executive Survey 2016: An Update on the Adoption of Big Data in the Fortune 1000* Boston: New Vantage Partners LLC Accessed on http://newvantage.com/wp-content/uploads/2016/01/Big-Data-Executive-Survey-2016-Findings-FINAL.pdf
32. O'Leary, D.E., (2013) 'Big Data', The 'Internet of Things' and The 'Internet of Signs' *Intelligent Systems in Accounting, Finance and Management* Vol. 20, Pp53–65.
33. Park, H. W., and Leydesdorff, L., (2013) "Decomposing Social and Semantic Networks in the Emerging Science of "Big Data Science" *Journal of Informetrics* Vol. 7 No. 3Pp756- 766
34. Sharma, S., (undated) "An Extended Classification and Comparison of NoSQL Big Data. Models" Center for Survey Statistics and Methodology, Iowa State University, Ames, Iowa, USA.
35. Sajithra K, and Patil, R., (2013) "Social Media – History and Components" *Journal of Business and Management* Vol.7, Issue 1 Pp 69-74 Shelton, T., Poorthuis, A., Graham, M., Zook, M., (2014) "Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data'" *Geoforum* Vol. 52, Pp167–179.
36. Shi, G., Zhu, Y., Mi, S., Ma, J., and Wan, J., (2014) "A Big Data Mining in Petroleum. Exploration and Development" *Advances in Petroleum Exploration and Development* Vol. 7, No. 2, Pp1-8.
37. Smith, M.A., Shneiderman, B., Milic-Frayling, N., Rodrigues, E.M., Barash, V., Dunne, C.,

Capone, T., Perer, A., Mendes, E., and Gleave, E., (2009) *Analyzing (Social Media) Networks with Node XL* University Park, Pennsylvania, USA.
38. Stenberg, J., (1997) *Global Production Management in Newspaper Production and Distribution–Coordination of Products, Processes and Resources* Doctoral thesis submitted to Department of Manufacturing Systems Division of Graphic Arts Technology The Royal Institute of Technology Stockholm, Sweden.
39. Stephen, I., and Smith, M.M., (2013)"Birds of a Feather Tweet Together: Integrating Network and Content Analyses to Examine Cross-Ideology Exposure on Twitter" *Journal of Computer-Mediated Communication* Volume 18, Issue 2, Pp 40-60.
40. Thelwall, M., Sud, P., and Vis, F., (2011) "Commenting on YouTube Videos: From Guatemalan .
41. Rockto El Big Bang" *Journal of the American Society for Information Science and Technology* Pp1-19.
42. Willis, A., Fisher, A., and Lvov, I., (2015) "Mapping networks of influence: Tracking Twitter Conversations through time and space" *Journal of Audience and Reception Studies* Vol. 12, Issue 1 pp494-530.
43. Xiang, Z., Schwartz, Z., and Gerdes, J.H., and Uysal, M., (2015) "What can big data and text
44. Analytics tell us about hotel guest experience and satisfaction?" *International Journal of Hospitality Management* Vol. 44, Pp120–130.
45. Yang, M., Kiang, M., and Shang, W., (2015) "Filtering big data from social media – Building an Early warning system for adverse drug reactions" *Journal of Biomedical Informatics* Vol. 54, Pp230–240.
46. Yasseri, T., Hale, S.A., and Margetts, H.Z., (2014) Modeling the rise in Internet-based petitions.
47. Yu, Y., and Wang, X. (2015) "World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fans' tweets" *Computers in Human Behavior* Vol.48, Pp392–400. doi: 10.1016/j.chb.2015.01.075.
48. Zhou, X., and Xu, C., (2017) "Tracing the Spatial-Temporal Evolution of Events Based on Social Media Data" *International Journal of Geo-Information* Pp1-15.

Appendix

Code Sheet- Frequency of Categories Used

Focus					
Business	Politics	Economy	Relationship	News	Others
5	6	0	13	2	5

Methods				
Exploratory	Survey	Computational	Content Analysis	Others
7	4	13	8	5

Theories		
Network	Graph	Others
6	2	12

Data Collection and Analysis					
NodeXL	NoSQL	MySQL	Hadoop	MongoDB	Others
1		1	2	1	24

Sampling	
Probability	Non-Probability
6	23

6/15/2019