Vol 8. No. 3 Issue 2 – October, 2015
African Journal of Computing & ICT

Mining Social Media for Conflict Prevention and Resolution

K.P. Mensah
Department of Computer Science
University for Development Studies
Navrongo, UER-Ghana.
menkpat@yahoo.com /pmensah@uds.edu.gh.

S. Akobre
Department of Computer Science
University for Development Studies
Navrongo, UER-Ghana. aweyass@yahoo.com
School of Computer Science and Engineering
University of Electronic Science and Technology of China
4Section2,North Jianshe Road Chengdu,Sichuan, P.R.China

ABSTRACT

The power of social media such as Twitter, Facebook, Instagram, LinkedIn, etc. in our daily lives cannot be underestimated. Governments have been toppled and countries destabilized as a result of sentiments expressed by citizens on social media. In this paper, we show that mining Twitter Follower/Friend network structure and data can be a powerful method to recognize the needs, sentiments, opinions and interests of the citizenry. Hierarchical clustering and Partition around Medoids were used. It was discovered that the Twitter community in Ghana takes delight in discussing political parties and personalities instead of pressing issues like corruption and unemployment. Follower/Friends network analysis was used to discover influential “e-people” who could serve as potential mediators during conflict situations. This method is aimed at identifying the most influential people in the Ghanaian Twitter Community and to discover what most people are complaining about through their tweets. This can be used to avoid a replication of the “Arab Spring” elsewhere. Possible Mediators can also be discovered. We propose an inexpensive but effective method to help prevent and resolve the rampant conflicts in the World that arise due to neglect of citizens by their governments. Advertisers, policy makers and political parties also stand to benefit from this approach.

Keywords: Social Media, Data Mining, Conflict Resolution, Social Network, Betweeness, Centrality, Eigen Vector.

African Journal of Computing & ICT Reference Format:
K.P. Mensah & S. Akobre (2015): Mining Social Media for Conflict Prevention and Resolution.
Afri J Comp & ICTs Vol 8, No.3 Issue 2 Pp 163-170.

1. INTRODUCTION

Most countries in the world; especially third world countries, are vulnerable to wars due to poverty, unemployment, corruption and bad governance. It has been established that about 60% of the population in many of these countries are made up of the youth [1] who are largely unemployed and very active on social media [2]. The Arab Spring [3], [4] in North Africa was fueled by the use of social media which has left in its wake conflicts in the region. As most African countries embrace democracy, it is common occurrence for tensions to rise during elections. We propose the use of unsupervised clustering algorithms and social network analysis to mine Twitter with the aim of identifying issues that could be potential starting point(s) for conflict during elections. Our method can also identify users who could serve as possible mediators during conflict situations. Ghana is used as a case study since 2016 is an election year **in** the country. Even though majority of social media users in Ghana are on Facebook, the middle class and decision makers prefer to use Twitter to express their opinions on important issues; hence our decision to choose Twitter for analysis.

1.1 Clustering

Clustering is a machine learning technique that group data based on their similarity. Clusters that are formed are distinct and no data point is categorized in more than one cluster. The method is widely implemented as unsupervised. Unsupervised clustering does not need training/test data when making clusters as the algorithm is guaranteed to discover the relationship between objects.

The objective is to maximize the similarity between data points of the same cluster while at the same time minimizing the similarities between data points of different clusters. Algorithms such as K-Means [5] and Hierarchical clustering are implemented as unsupervised algorithms. Distances between objects in a cluster, between objects and other clusters, and between clusters are very important for correct placement of an object in the appropriate partition. Several measures can be used to calculate the distances between objects in a cluster.

A good distance measure is a function *f(x,y)* that takes two data points *x* and *y* such that all the following conditions are satisfied:

1. Symmetry: $f(x,y) = f(y,x)$.
2. Equality: i $x = y$ then $f(x,y) = 0$.
3. Triangular inequality symbolizes the shortest path property of clustering distance measures and is given by:
   $f(x,y) \le f(x,z) + f(z,y)$.
4. No negative distances $f(x,y) > 0$.

The following distance measures are commonly used in clustering algorithms:

1. *Manhattan distance*: It takes the absolute difference of the distances between objects x and y
   $$D_m(x,y) == \sum_{i=1}^{n} |X_i - Y_j| \quad (1)$$
2. *Euclidea distance*: It evaluates the distances of alternate paths between given objects in a cluster and takes the path with the shortest distance. It is the most widely used distance metric for clustering. Assuming the objects are *x* and *y* at distance D apart, then
   $$(x,y) = \sqrt{\sum_{i=1}^{n} (X_i - Y_j)^2} \quad (2)$$
3. *Minkowski distance*: It Generalizes the *Euclidean distance* to provide some flexibility in choosing the parameter p, which is 2 in the *Euclidean distance*. The expression for the *Euclidean distance* can be re-written as: $D_e = \left(\sum_{i=1}^{n} |X_i - Y_j|^2\right)^{1/2}$
   (3)
   Generally,
   $$D_{min} = \left(\sum_{i=1}^{n} |X_i - Y_j|^p\right)^{1/p}. \quad (4)$$
4. *Cosine similarity measure*: It measures the cosine of the angle between the two objects/vectors with integer/boolean components. It is widely used when clustering transactional data.

In hierarchical clusters, the distances between clusters can be determined either by finding the distance between the nearest points in the two clusters called *single linkage*, or the distance between the farthest points in the cluster called *complete linkage*, or *average linkage* which is the average distance between all the points in the cluster. Clustering algorithms have their own internal mechanisms used to evaluate performance. For instance, the k-means clustering algorithm finds the squared distance of the data point to the cluster center (sum of squares error) to determine how acceptable a cluster is. In addition, external methods of evaluation could be employed. A separate set of data could be used to measure how representative the clusters are. Parameters such as F-measure, purity, entropy, and random index can then be calculated. A delicate but difficult issue during cluster generation is determining the appropriate number of clusters [6] for the dataset.

Proposed solutions [7] to this problem include *rule of thumb*; i.e. finding the square root of half the number of objects *(dp)* as a rough estimate to the number of clusters *(c)*: $c = \sqrt{dp/2}$. Other methods include E*lbow Method*, *Cross-Validation Method, Silhouette Method* and the *Aligned Box Criteria (ABC)* [8].

1.2 Social Network Analysis

Social Networks are depicted as graphs. They are made up of entities that may (may not) share common characteristics in a given locality. They have been applied in Sociology [9] long before social media giants such as Facebook, Twitter, etc. immerged. These Social Networks are a repository of vast amounts of data; mining of which could result in unearthing relationships that could impact real life situations. Aside this, the structure of the social graph could also be mined. For instance, the concept of centrality is used to determine how important a given individual (node) is in the network. Degree centrality of a directed network can be computed for both in-degree and out-degree. For an undirected graph, the degree centrality d_c of the *jth* node v_j is given by the number of edges d_j adjacent to v_j; $d_c(v_j) = d_j$. This measure indicates how popular an individual is in a network; that is the higher the degree, the popular the individual. However, it is not entirely accurate to use degree centrality to measure "social status" in a network, since it is not all the connections that link to important nodes. To take into consideration the status of the node(s) to which v_i is connected in the network, we use Eigen vector [10] centrality e_c to generalize the degree centrality measure:

$$e_c(v_i) = \frac{1}{\lambda} \sum_{j=1}^{n} A_{j,i}\, e_c(v_j) \quad (5)$$

where λ = some eigenvalue (a constant), and $A_{j,i}$ = adjacency matrix. Letting E_c be the *n*x*1* matrix (transpose) of the above quantity, we can rewrite it as $E_c = A^T E_c$. Notice that for an undirected network, *A* is the same as A^T. E_c tells us which edges the individual is likely to be using after a long time. The Perron-Frobenius theorem [11] is used to avoid negative Eigenvector centrality values. Katz Centrality measure [11] avoids the limitation of Eigenvector centrality that occurs in directed acyclic graphs. It introduces the parameter β to prevent zero centrality values:

$$e_k(v_i) = \alpha \sum_{j=1}^{n} A_{j,i}\, e_k(v_j) + \beta. \quad (6)$$

A limitation of the Katz Centrality measure is solved by PageRank, which does not permit a central node to pass its importance to adjacent nodes. To share the centrality to each adjacent edge, PageRank divides the centrality among the outgoing edges of the node in a directed network:

$$P_r(v_i) = \alpha \sum_{j=1}^{n} A_{j,i} \frac{P_r(v_j)}{d_j^{out}} + \beta \quad (7)$$

The idea is that, connections will converge at the PageRank centrality node if nodes in the network were chosen randomly, and random out links were followed. Notwithstanding, PageRank has to contend with spider traps and dead ends [12]. To consider the importance of a node in a network, we can consider how often it is used as a bridge on the shortest path to connect other nodes; referred to as Betweeness centrality:

$$B_c(v_i) = \sum_{s \neq p \neq v_i} \frac{\sigma_{sp}(v_i)}{\sigma_{sp}} \qquad (8)$$

where σ_{sp} = the number of shortest paths from node s to p, and $\sigma_{sp}(v_i)$ = the previous quantity but for the ones that pass through node v_i. When v_i is on all the shortest paths between a and p, B_c assumes its maximum value. *Betweeness* centrality can be computed using Dijkstra's algorithm or Brandes' algorithm [13]. These centrality measures could be applied to a group of vertices.

Clustering can be used to find communities in a social network. In defining a distance measure for social network clustering, triangular nodes should be taken into account. Using the k-Means clustering algorithm, a network could be clustered into communities [12]. A node will only be assigned to a cluster if it has the shortest average distance to all the other nodes in the cluster.

2. RELATED WORKS

Paul et al. [14] proposed a probabilistic topic model called the Ailment Topic Aspect Model (ATAM) used to monitor the spread of ailments that are discussed on twitter. The model was able to group symptoms and treatments for ailments into the appropriate public health related topics. Johansson et al. [15] describes a semi-automatic system involving the automatic harvesting of online data from humanitarian organizations' reports, Twitter, Facebook and Blogs to forecast where the next conflict will be and on what issue. Park et al. [16] analyzed depressive moods of users portrayed in tweets.

They concluded that users who tweeted depressive sentiments were actually depressed, and that social media could be an important source of data for clinical studies. In [17], the researchers used indegree, retweets, and mentions to study the influence of a user on twitter. The behavior of social network users was characterized in [18]. [19] Used frequent sets in association rules to predict the outcome of events. Ediger et al. [20] proposes GraphCT; a Graph Characterization Toolkit used to represent social network graph data. It has the ability to determine the network centrality measures on large datasets. There are numerous literature [21], [22], [23], [30] involving the mining of social media data such as tweets for one reason or the other. This paper adds to the works already mentioned by mining both social media data and its network structure for conflict prevention and resolution.

Table 1 presents the list of search terms used to collect tweets. This last sample was made up of 8,951 tweets. It was collected between 3rd to 11th September 2015 and used for clustering.

Table 1: A List of Search terms used to collect tweets

Term	Average Number of tweets
npp ghana	447
new patriotic party ghana	37
new patriotic party	259
npp	560
ndc	3000
ndc ghana	300
corruption in ghana	350
unemployment in ghana	250
national democratic congress	700
national democratic congress ghana	7
ghana politics	1084
politics in ghana	260
political parties in ghana	8
ghana political parties	11
John mahama	1023
president john dramani mahama	229
president jdm	29
jdm	186
akuffo addo	142
nana akuffo addo	41
ruling party Ghana	3
opposition party Ghana	25
Total	8,951

3. METHODS

3.1 Data Description

R's twitteR [24] package was used to collect tweets based on the search terms in Table 1. This was carried out every week from 3rd September 2014 to 11th September, 2015. The same search terms were used and purposely chosen in the Ghanaian context. A minimum of 8,120 tweets were collected each time within the said period. They were made up of original tweets, retweets and replies. Figure 1 depicts, on the average, the number of re-tweets each of the most influential users obtained throughout the period.

Over the 32 weeks that data was collected, the eight users in Figure 1 were found to be the most influential users based on retweet count with a statistic of 0.90. Based on 95% confidence interval, a 0.10 margin of error was obtained.

Vol 8. No. 3 Issue 2 – October, 2015
African Journal of Computing & ICT

www.ajocict.net

This indicates that the probability that the eight users chosen are the most retweeted for all the 32 samples is between 0.80 and 1.00. The level-1 Friends and Followers of the eight most retweeted users were collected. A total of 1062 Friends and Followers were obtained for the eight users. The network diagram of Friends/Followers using re-tweet count as a measure of influence on twitter [17] is shown in Fig. 2. Figure 3 shows the degree distribution of the network on a Log-Log scale. The average degree was 1.028.

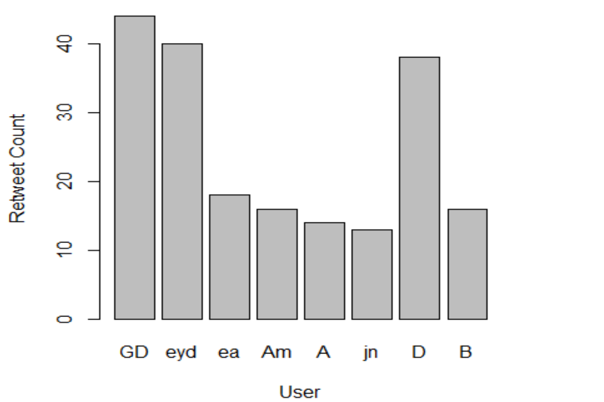

Fig. 1 Re-tweet count showing how many of each of the eight users' tweets were re-tweeted by others

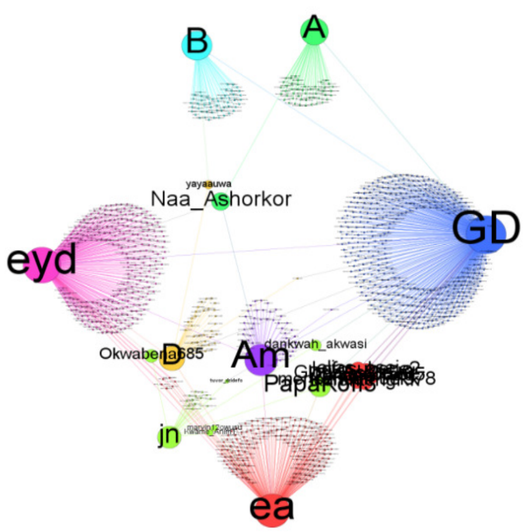

Fig. 2 Force Atlas Layout algorithm applied on Friends and Followers network

3.2 Data Cleaning

In order not to produce trivial results, well known political figures, celebrities and news outlets such as *Citi973*, *newsontv3*, etc. were ignored in the retweet count leading to the choice of the eight most retweeted users.

Re-tweets were not ignored during clustering. It is assumed that a user who re-tweets another user's tweets would have tweeted same if the idea had come to them first. Screen names of the most re-tweeted users were masked for anonymity. Users whose tweets were most replied-to were also obtained, however it is difficult to interpret because another user may reply to a tweet to show their approval or disapproval. Unwanted characters such as punctuations, tabs and numbers were replaced with spaces and the corpus stemmed.

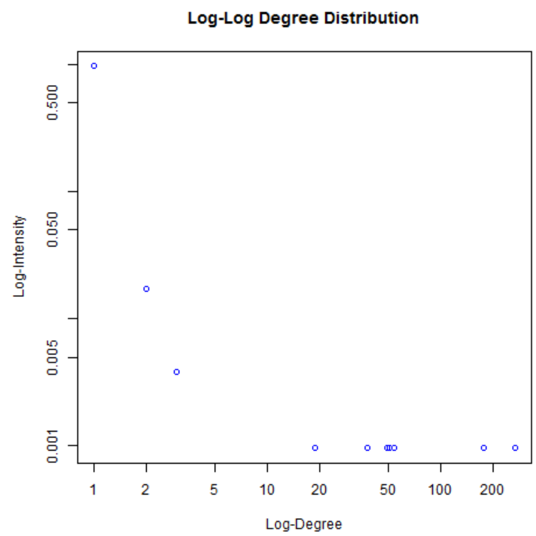

Fig. 3 Degree distribution of the network

Hyperlinks and references were removed from the corpus. Stop words such as *the*, etc. were also removed from the corpus.

4. DISCUSSION OF RESULTS

R's igraph [25] package was used to build a graph of Friends and Followers of the eight most retweeted users. The network has 1039 nodes and 1068 edges. Since the depth of the graph from the most re-tweeted users to their Followers/Friends is one, links to both friends and followers can be treated as undirected links. In addition, our interest is in finding the centrality measures of the eight most re-tweeted users. Fig. 2 depicts the resulting network with the Force Atlas Layout algorithm. This algorithm allows linked nodes to attract each other than non-linked nodes. Fig. 4 shows the degree centrality for the network under consideration. The user represented with the largest node in the network has the highest number of edges. This could be explained with the preferential attachment model [26]; i.e. the probability of a new user following an existing user is proportional to the number of followers the existing user already have.

Vol 8. No. 3 Issue 2 – October, 2015
African Journal of Computing & ICT

www.ajocict.net

According to [17], in-degree is not a good measure of influence; however it can be used as a measure of how rapid a node can diffuse information. Based on Clustering Coefficient and the transitive nature of following; (i.e. the "Followers of my Followers" are also my Followers), user GD in Fig. 4 will be a good starting point for information propagation in the network beyond what is shown here. Fig. 5 shows how important some nodes are in serving as the shortest paths between nodes on the network. We notice that, user eyd has high betweeness centrality than GD who has high degree centrality from Fig. 4. Assuming we were considering a business network, eyd's position in Fig. 5 would be that of a broker.

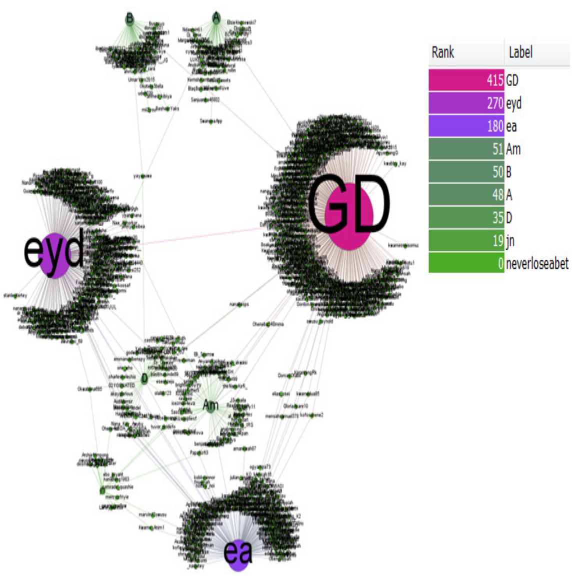

Fig. 4 Degree centrality shows user *GD* has the highest degree

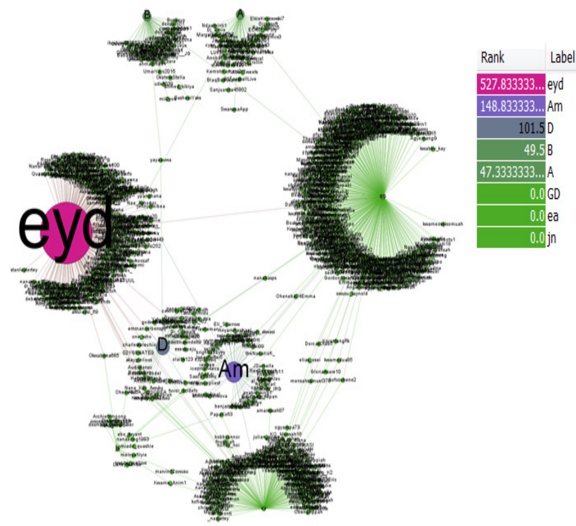

Fig. 5 Betweeness centrality shows *eyd* having a high betweeness.

As a political network, *eyd* can serve as a mediator between the different communities that uses it as a bridge. *eyd* also serves as a better medium for information diffusion across sub-networks due to its betweeness. PageRank centrality (Fig. 6) shows that a random surfer on this network will spend a large fraction of time on user *Okwabena685* than any other user. This means that, *Okwabena685* is the appropriate user who is best suited to preach peace to any new user who has not already taken sides in the network during conflict. Hierarchical clustering was used to cluster the network. Eight communities were detected as shown in Figures 7.

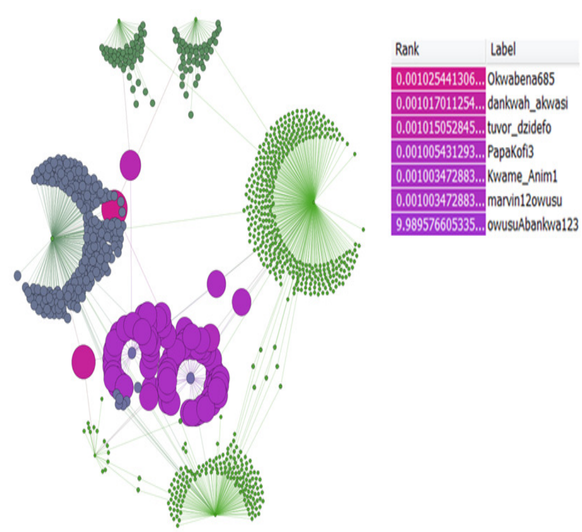

Fig. 6 User *Okwabena685* has the highest ranking in terms of Page-rank centrality

Vol 8. No. 3 Issue 2 – October, 2015
African Journal of Computing & ICT

www.ajocict.net

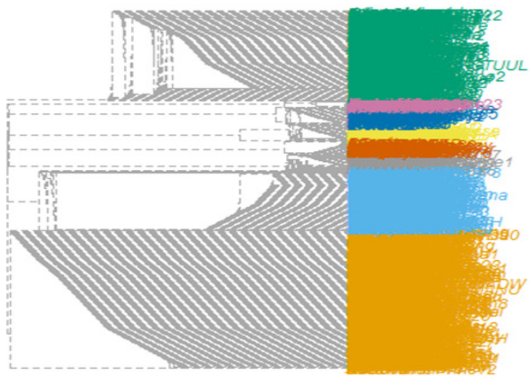

Fig. 7 Clustering of Friends and Followers network detected eight Communities

4.1 Clustering of Words and Tweets

The dataset was cleaned using the *tm* package [27] in R. In order to visualize the most important words in the corpus, a word cloud of words of frequency not less than 1000 was generated with its associated bar chart as shown in Fig. 8. Hierarchical clustering using three cluster centers (k=3) was applied to cluster the words forming the Tweets. It can be seen from Fig. 9 that "*ndc*" being the ruling party is in its own cluster. The current "*president*", "*john*" "*mahama*" are in the same cluster, whilst the opposition "*npp*", its flag-bearer "*nana*" "*addo*" and the newly elected Nigerian president "*buhari*" are in the same cluster. The last cluster is intuitive because when president "*buhari*" was elected, people started using his age to justify why "*nana*" "*addo*" could still be a president despite his age. During the period data was being collected for this work, changing the Ghanaian voter's register and corruptions in the judiciary were hot topics under discussion. However, none of these issue-based topics featured in Figures 8, 9and 10. This may suggest that Ghanaians are more interested in discussing political parties and personalities rather than issues.

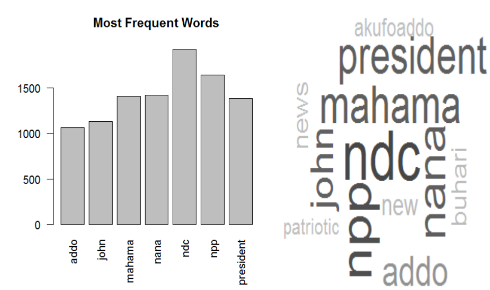

Fig 8 Bar plot and Word cloud of the corpus shows *ndc* and *npp* (both political parties) as the most important/frequent words.

The tweets were clustered using Partition Around Medoids (PAM) with the Euclidean distance metric. The aim was to determine if group of tweets were discussing the same topic or person. PAM; a form of k-Medoids algorithm was chosen for the clustering due to the fact that it is robust to noise such as outliers.

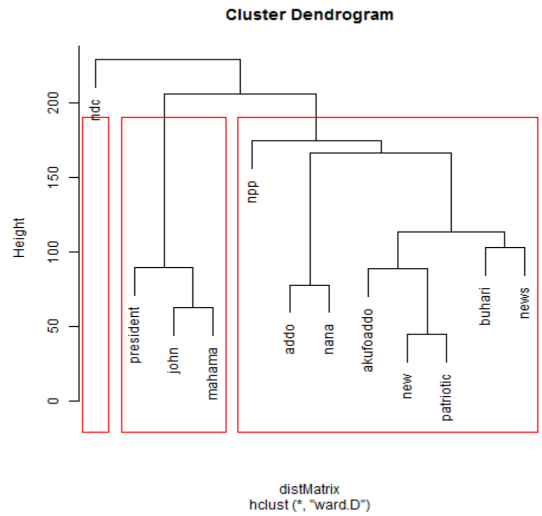

Fig. 9 Hierarchical clustering of words using 3-cluster centers shows that personalities and political parties are discussed the most

PAM cluster centers are represented by objects (Medoids) closer to the center of the cluster instead of Means as in k-Means algorithm. Specifically, a variant of PAM called PAMK [28] was used since it does not have the limitation of letting the user choose the number of clusters. Fig. 10 is a 2-dimensional cluster plot of applying PAMK on the corpus. 10 clusters were generated. An average silhouette width of 0.55 was obtained suggesting that the partitions obtained by the clusters are separated from one another. Particularly, clusters 6, and 8 were well separated as shown by their silhouettes. However, cluster 4 overlaps all other clusters and tweets belonging to this cluster could not fit well into the other clusters. Clusters 1, 2, 3 and 7 contains tweets on "*npp*", its opposition leader "*addo*" "*nana*", etc. The rest of the clusters were centered on "*john*" "*mahama*" and "*buhari*". These clusters also confirm the assertion that Ghanaians discuss political parties and personalities instead of core issues of "*bread and butter*".

4.2 Limitations

Data was collected from Twitter using the Ghanaian political environment as a case study. However, Twitter is not so popular with the ordinary internet user in Ghana like Facebook [29].

Vol 8. No. 3 Issue 2 – October, 2015
African Journal of Computing & ICT

www.ajocict.net

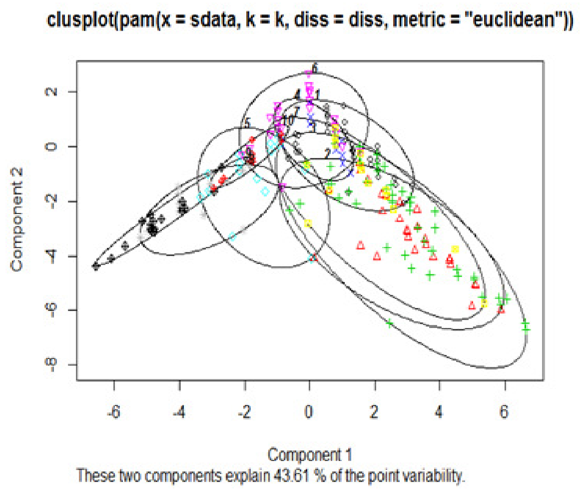

Fig. 10 PAMK cluster plot of tweets using 10 cluster centers confirms the conclusion drawn by Fig.8

As at the time data was collected, Twitter has a limit to the number of words a user can use to express their opinion on a subject. As a result, jargons and characters can be used to express valuable information in tweets. It is also not everyone who has access to the internet. Despite these limitations, this work has shown that it is possible to obtain valuable knowledge from social media to enable policy-makers act before things go out of control.

5. CONCLUSIONS AND FUTURE WORK

This paper has demonstrated that the structure of social media can be mined to identify influential people who could serve as mediators or information propagators during conflict situations to avoid a repeat of the "Arab Spring" elsewhere. Advertisers can take advantage of the methods outlined in this paper to enable their product information reach wide audiences. Application of unsupervised clustering algorithms revealed that people, rather than issues are mostly discussed in Ghanaian politics; meaning that if elections were to be held in Ghana today, people may not vote based on issues but instead on personalities and party affinities.
As future work, it will be desirable to fully automate the methods outlined in this paper.

REFERENCES

[1] Adebayo, A. A. (2013), *Youths' unemployment and crime in Nigeria: A nexus and implications for national development*, International Journal of Sociology and Anthropology, Vol. 5(8), pp. 350-357.

[2] Livingstone, S. (2008), *Taking risky opportunities in youthful content creation: teenagers' use of social networking sites for intimacy, privacy and self-expression*. New media & society, 10 (3). pp. 393-411.

[3] Wolfsfeld, G., Segev, E. and Sheafer, T. (2013), *Social Media and the Arab Spring: Politics Comes First*. The International Journal of Press/Politics, XX(X), pp. 1–23.

[4] Storck, M. (2011), *The Role of Social Media in Political Mobilisation: a Case Study of the January 2011 Egyptian Uprising*, M.A. Thesis: University of St Andrews, Scotland.

[5] Hartigan, J. A. and Wong, M. A. (1979), *A K-Means Clustering Algorithm*, Journal of the Royal Statistical Society. Vol. 28, No. 1, pp. 100-108.

[6] Pham, D. T., Dimov, S. S. and Nguyen, C. D. (2004), *Selection of K in K-means clustering*, Proc. IMechE Vol. 219 Part C: J. Mechanical Engineering Science, pp.103-119.

[7] Bell, J. (2015), *Machine Learning: Hands-On for Developers and Technical Professionals*, John Wiley & Sons, Inc., ISBN: 978-1-118-88939-8 (ebk), pp. 167-168.

[8] Dean, J. (2014), *Big Data, Data Mining, and Machine Learning*, John Wiley & Sons, Inc., ISBN 978-1-118-92069-5 (ebk), pp. 135-137.

[9] Bonchi, F., Castillo, C., Gionis, A., and Jaimes, A. 2011. *Social network analysis and mining for business applications*. ACM Trans. Intell. Syst. Technol. 2, 3, Article 22 (April 2011), 37 pages.

[10] Sterling, M. J., (2009), *Linear Algebra for Dummies*, Wiley Publishing, Inc., ISBN: 978-0- 470-43090-3, pp. 289.

[11] Zafarani, R., Abbasi, M.A. and Liu, H., (2014), *Social Media Mining*, Cambridge University Press, Draft Version, pp. 76-78.

[12] Leskovec, J., Rajaraman, A. and Ullman, J.D. (2014), *Mining of Massive Datasets*, Stanford University, pp. 163-170.

[13] Brandes, K., (2001), *A faster algorithm for betweenness centrality*, Journal of Mathematical Sociology 25 (2001), no. 2, pp. 163–177.

[14] Paul, M.J. and Dredze, M., *A Model for Mining Public Health Topics from Twitter*, [Online URL: www.cs.jhu.edu/~mpaul/files/2011.tech.twitter_health.pdf] [accessed 09/10/2015].

[15] Johansson, F., Brynielsson, J., Horling, P., Malm, M., Martenson, C., Truve, S. and Rosell, M., *Detecting Emergent Conflicts through Web Mining and Visualization* [Online URL: https://www.recordedfuture.com/assets/Detecting-Emergent-Conflicts-through-Web-Mining-and-Visualization.pdf] [accessed 07/10/2015].

[16] Park M, Cha C, Cha M, *Depressive Moods of Users Portrayed in Twitter*. HI-KDD '12, Beijing (2012).

[17] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. (2010), *Measuring user influence in twitter: The million follower fallacy*, In 4th international aaai conference on weblogs and social media (icwsm), Vol. 14, No. 1, pp. 8.

[18] Benevenuto, F., Rodrigues, T., Cha, M., and Almeida, V. (2009), *Characterizing user behavior in online social networks.* In Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference, pp. 49-62.

[19] Pavlyshenko, B., *Forecasting of Events by Tweet Data Mining*. [Online URL: http://arxiv.org/pdf/1310.3499][accessed 6/10/2015].

[20] Ediger, D., Jiang, K., Corley, C., Farber, R. and C Reynolds, W.N., (2010), *Massive Social Network Analysis: Mining Twitter for Social Good*, 39th International Conference on Parallel Processing, IEEE Computer Society, pp. 583-593.

[21] Godfrey, D., Johns, C., Sadek, C., Meyer, C. and Race, S., *A Case Study in Text Mining: Interpreting Twitter Data From World Cup Tweets*. [Online URL: http://arXiv:1408.5427v] [accessed 07/10/2015].

[22] Reips, U. and Garaizar, P., (2011), *Mining twitter: A source for psychological wisdom of the crowds*. Psychonomic Society, Inc. 2011, Behav Res., DOI 10.3758/s13428-011-0116-6.

[23] Ashktorab, Z., Brown, C., Nandi, M. and Culotta, A., (2014), *Tweedr: Mining Twitter to Inform Disaster Response*, Proceedings of the 11th International ISCRAM Conference–University Park, Pennsylvania, USA.

[24] Gentry, J. (2012), *twitteR: R based Twitter client*. R package version 0.99.19. [Online URL: http://cran.r-project.org/web/packages/twitteR/vignettes/twitteR.pdf [Accessed 20/09/2015].

[25] Csardi, G. and Nepusz, T. (2006), *The igraph software package for complex network resear ch.* InterJournal, Complex Systems: 1695.

[26] Barabasi, A. L. and Albert, R. (1999), *Emergence of scaling in random networks*. Sci. 286, 5439, pp. 509–512.

[27] Feinerer, I. (2012), *tm: Text Mining Package*. R package version 0.5-7.1. [Online URL: http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf [Accessed 12/10/2015].

[28] [28] Hennig, C. (2010), *fpc: Flexible procedures for clustering*. R package version 2.0-3. [Online URL: [Accessed 12/10/2015].

[29] [29] Mensah, K. P. (2015), *Textual Prediction of Attitudes towards Mental Health*, Int. J. Knowledge Engineering and Data Mining, Vol. 3, Nos. 3/4, 2015.

[30] [30] De Choudhury, M., Monroy-Hernández, A. and Mark, G. (2014), *"Narco" Emotions: Affect and Desensitization in Social Media during the Mexican Drug War*, CHI 2014, Toronto, ON, Canada, ACM 978-1-4503-2473-1/14/04.

Author's Biographies :

Mensah Kwabena Patrick is a Lecturer with the Department of Computer Science, Faculty of Mathematical Sciences at the University for Development Studies, Ghana, since 2012. He obtained his BSc in Mathematical Science (Computer Science option) in 2009 from the same institution. In 2011, he received his MSc in Computer Science from the African University of Science and Technology (AUST), Abuja. He is into Data Mining, Machine learning and Residue Number Systems.

Stephen Akobre received his Bsc. degree in Computer Science in 2006 and Msc. degree in Telecommunications Engineering in 2011 from the Kwame Nkrumah University of Science and Technology, Kumasi-Ghana. In 2007 he joined the University for Development studies as a research assistant. He is now a lecturer at the Department of Computer Science, University for Development Studies, Navrongo Campus. His research interest include effect of propagation impairments on satellite communications systems, data mining, big data and machine learning. He is currently pursuing his PhD in Computer Science and Technology at the University of Electronic Science and Technology of China. He is also a student member of the IEEE UESTC-Chengdu Chapter.