Twitter Analysis using Matlab TMG

Zarak Farid, Nouman Zeb, Sherjeel Sikander

[#]*Dept of Computer Science and Engineering, GIKI, Topi, KPK, Pakistan*
u2010357@giki.edu.pk, u2010277@giki.edu.pk, u2010337@giki.edu.pk

Data Mining & Warehousing – CS 437

Abstract: **In recent years, social media has become ubiquitous and important for social networking and content sharing. And yet, the content that is generated from these websites remains largely untapped. This paper illustrate the analysis done on four major events within a time span of three months considering tweets of 2-3 local journalists and 2-3 foreign journalists. Using K-Means Clustering and Ward clustering, four clusters are formed for each event and the top three results of each cluster are analysed to see what people are discussing in the duration of that event**.

***Keywords*— Twitter Analysis, Clustering, K-Means, Ward**

I. Introduction

Social Media has become a common mode of communication in the 21st century with people from all ethnicities, groups, communities and backgrounds sharing views and opinions on a common platform. Twitter is a social networking and microblogging service started in 2006 utilising instant message, SMS or a web interface where users can follow journalists, celebrities or new channels etc. to stay in touch with their favourite topics and to express their opinion. Because of its ease of use, speed and reach, social media is fast changing the public discourse in society and setting trends and agendas in topics that range from the environment and politics to technology and the entertainment industry.

The report utilizes the twitter platform in conjunction with several events to analyse the most discussed topics during the duration of that event.

II. Related Work

Although Twitter has been very popular as a web service, there has not been considerable published research on it. Huberman and others [1] studied the social interactions on Twitter to reveal that the driving process for usage is a sparse hidden network underlying the friends and followers, while most of the links represent meaningless interactions. Java et al [4] investigated community structure and isolated different types of user intentions on Twitter. Jansen and others [2] have examined Twitter as a mechanism for word-of-mouth advertising, and considered particular brands and products while examining the structure of the postings and the change in sentiments.

There has been some prior work on analyzing the correlation between blog and review mentions and performance. Gruhl and others [6] showed how to generate automated queries for mining blogs in order to predict spikes in book sales. And while there has been research on predicting movie sales, almost all of them have used meta-data information on the movies themselves to perform the forecasting, such as the movies genre, MPAA rating, running time, release date, the number of screens on which the movie debuted, and the presence of particular actors or actresses in the cast. Joshi and others [7] use linear regression from text and metadata features to predict earnings for movies. Sharda and Delen [5] have treated the prediction problem as a classification problem and used neural networks to classify movies into categories ranging from 'flop' to 'blockbuster'. Zhang and Skiena [3] have used a news aggregation model along with IMDB data to predict movie box-office numbers.

Our work however focuses more on the current discussion of local and foreign journalists during the given events.

III. Clustering Algorithms

Cluster analysis involves sorting data objects (or items) into natural groupings based on similarity. Grouping data is important because it can reveal information about the data such as outliers, dimensionality, or previously unnoticed interesting relationships. In cluster analysis there is often no prior specification about the number or nature of the groups to which the objects will be assigned. The grouping is often done based solely on similarity measures, and the ideal number of groups is often determined within the clustering algorithm. There are several main classes of methods in cluster analysis, including hierarchical clustering, partitional clustering, and model-based clustering.

The results in our case are based on two clustering algorithms used hierarchical method (Ward) and partitioning method (K-means).

A. Ward Linkage

Ward Linkage is hierarchical clustering method. While Ward's method is similar to the linkage methods in that it begins with N clusters, each containing one object, it differs in that it does not use cluster distances to group objects. Instead, the total within-cluster sum of squares (SSE) is computed to determine the next two groups merged at each step of the algorithm. The error sum of squares (SSE) is defined (for multivariate data) as:

$$SSE = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})^2$$

Where y_{ij} is the j_{th} object in the i_{th} cluster and n_i is the number of objects in the i_{th} cluster.

B. K Means

K-Means is a partitioning clustering method. MacQueen (1967) introduced the k-means method as an alternative to hierarchical clustering methods. This method is more efficient than hierarchical clustering, especially for large data sets and high-dimensional data sets.
The basic algorithm for the k-means method is as follows:

1. Specify the number of clusters k and then randomly select k observations to initially represent the k cluster centers. Each observation is assigned to the cluster corresponding to the closest of these randomly selected objects to form k clusters.

2. The multivariate means (or "centroids") of the clusters are calculated, and each observation is reassigned (based on the new means) to the cluster whose mean is closest to it to form k new clusters.

3. Repeat step 2, until the algorithm stops when the means of the clusters are constant from one iteration to the next.

In the traditional k-means approach, "closeness" to the cluster centers is defined in terms of squared Euclidean distance, defined by:

$$d_E^2(\mathbf{x}, \bar{\boldsymbol{x}}_{\boldsymbol{c}}) = (\mathbf{x} - \bar{\boldsymbol{x}}_{\boldsymbol{c}})'(\mathbf{x} - \bar{\boldsymbol{x}}_{\boldsymbol{c}}) = \sum_k (x_{ik} - \bar{x}_{ck})^2,$$

Where $x = (x_1, ..., x_p)'$ is any particular observation and $\bar{x}_c$ is the centroid for, say, cluster c.

IV. Methodology

The general flow of the algorithm and flow of data is as follows:

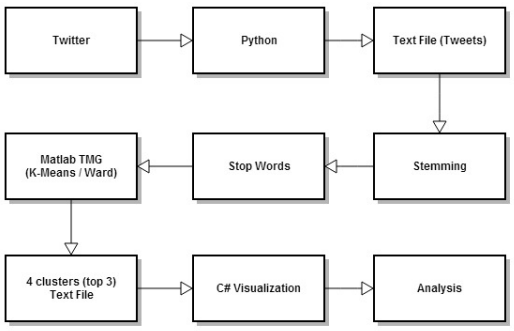


Figure 1 - General Flow of Twitter Analysis

1. Python selects the journalist tweets for a given date range and saves it in a text file.
2. The text file is loaded into Matlab and Stemming is applied on the data set.
3. The data words are reduced by removing the stop words using a predefined stop word list.
4. The data matrix is formed with rows as tweets and columns as words.
5. K-Means / Wards clustering is applied using the TMG Matlab library using 4 clusters and top 3 results are then stored in a text file.
6. Step 1-5 are repeated for each of the 4 events (each event further divided into 2 parts) and results are saved.
7. The collected top 3 results are then fed into C# application for visualization in a time line format.

V. Information Retrieval and Cleansing

Information Retrieval (IR) is essentially a matter of deciding which documents in a collection should be retrieved to satisfy a user's need for information. The user's information need is represented by a *query* or *profile*, and contains one or more *search terms*, plus perhaps some additional information such importance weights. Hence, the retrieval decision is made by comparing the terms of the query with the *index terms* (important words or phrases) appearing in the document itself. The decision may be binary (retrieve/reject), or it may involve estimating the degree of relevance that the document has to the query.

Unfortunately, the words that appear in documents and in queries often have many morphological variants. Thus, pairs of terms such as "computing" and "computation" will not be recognized as equivalent without some form of natural language processing (NLP).

A. Stemming

In most cases, morphological variants of words have similar semantic interpretations and can be considered as equivalent for the purpose of IR applications. For this reason, a number of so-called *stemming Algorithms*, or *stemmers*, have been developed, which attempt to reduce a word to its *stem* or root form e.g. run, runs and runner are considered as same word.

B. Limitations of Suffix Stemming Algorithm

The stemming algorithm used in TMG is suffix stemming. Suffix Stemming algorithms are sometimes regarded as crude given the poor performance when dealing with exceptional relations (like 'ran' and 'run'). The solutions produced by suffix stripping algorithms are limited to those lexical categories which have well known suffixes with few exceptions. Hence 'responsible', 'response', 'responsibility' will produce the same suffix 'respons', deleting the 'e' at the end as a result.

C. Stop Words

In computing, a stop word is a commonly used word (such as "the") that has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. When building the index, most engines are programmed to remove certain words from any index entry. The list of words that are not to be added is called a stop list. Stop words are deemed irrelevant for searching purposes because they occur frequently in the language for

which the indexing engine has been tuned. For twitter analysis these will outnumber any other word due to their frequent usage but provide little to no information as a result. Hence to produce more accurate analysis, these words are dropped and ignored before clustering.

VI. Events

The events focused for the analysis are summed up in the following table. The duration of the events is within 3 months to keep a timeline flow during visualization. The API is also restricted to 6 months of prior data to be gathered using Python library.

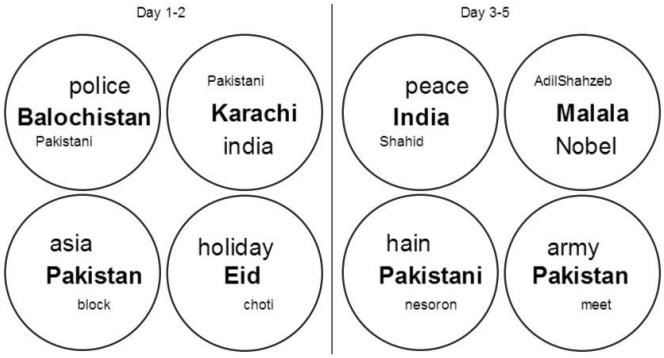
Sr #	Event	Date
1	Malala - Nobel Peace Prize	14th – 18th Oct 2013
2	Hakimullah killed in Drone	4th – 8th Nov 2013
3	Pindi attach - Muharram	18th – 22nd Nov 2013
4	Death Mandella	9th – 13th Dec 2013

Figure 2 - List of events selected for analysis

VII. Results

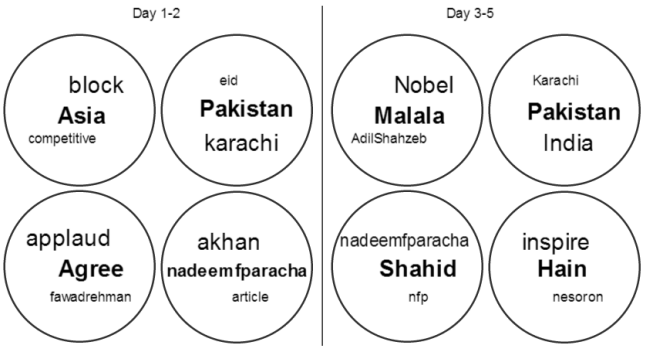
Each event was further broken down into 2 parts and K Means & Ward Linkage clustering algorithm was used. The results are visualized as follows.

A. *Malala – Nobel Peace Prize – K Means*



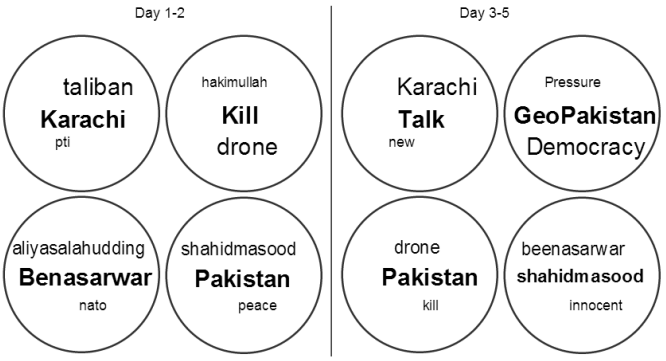
The nobel peace prize was discussed a couple of days after the event due to the eid holidays and balochistan politics which remained the more discussed topics at the beginning.

B. *Malala – Nobel Peace Prize – Ward Linkage*



The ward linkage shows similar results although a lot of meaningless noise data appears more frequent in a couple of the clusters.

C. *Hakimullah – Killed in Drone – K Means*

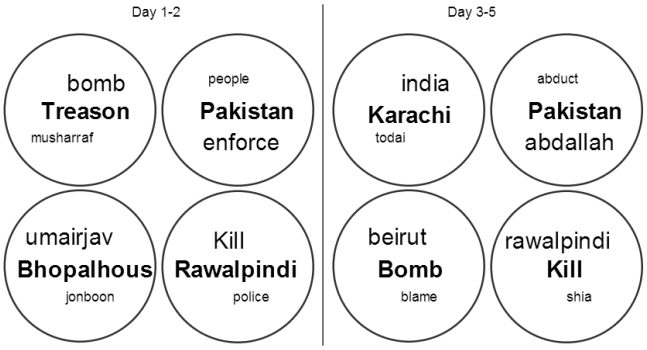


During the 5 day period hakimullah death remained the major point of discussion including Pakistan, peace and its security concerning terrorism.

D. *Hakimullah – Killed in Drone – Ward Linkage*

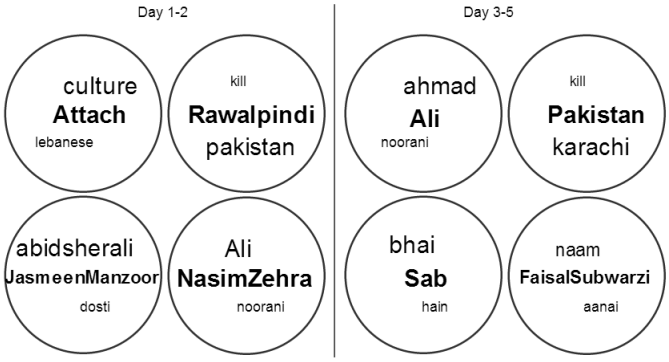
A lot of noise (Urdu words) appear to disturb the ward linkage analysis with the discussion of hakimullah only coming up in the 2nd part.

E. *Pindi Attack – Muharram – K Means*



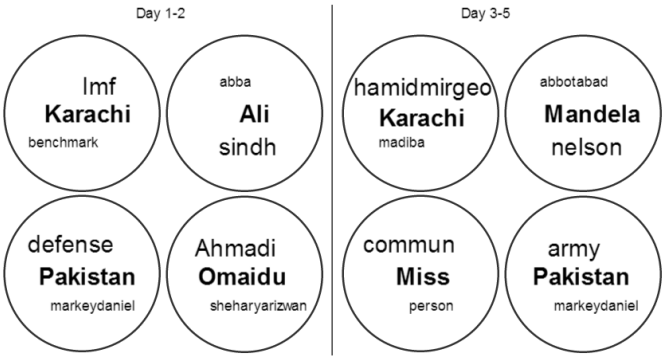
This duration seemed to have a lot of concurrent events including the Beirut Bombing and Amina Abdallah abduction hoax. Concerning the Rawalpindi killings the relevant discussion popping up is for the Pakistani government to enforce rule of law for the people and other security concerns. The people also discussed general politics regarding national treason and Musharraf. The 2nd part of the event shifted its focus a little from general rawalpindi killings to shia and sunni discussions with heated arguments.

F. *Pindi Attack – Muharram – Ward Linkage*



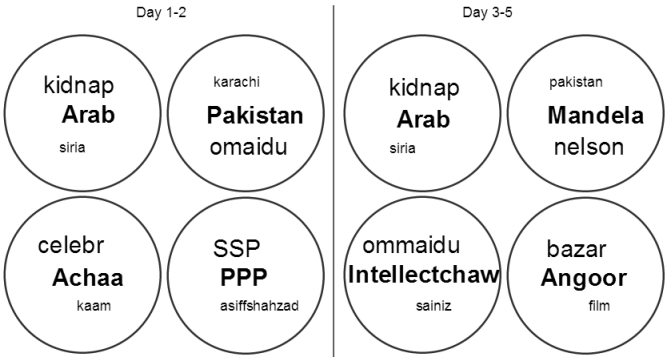
The ward linkage shows Rawalpindi killings with discussions fading a little in the other half. However the results are rather inaccurate (or less important) than the previous clustering.

G. *Death Madella – K Means*



The following clustering shows how people were less aware of Nelson Mandela at the time of his death as discussion grew only couple of days after his death. Religious talks regarding Ahmadi's also popped up with another major discussion of Pakistan and US Relations and how a war between them can be disastrous to both nations. People also discussed Supreme Court ordering of finding the missing persons.

H. *Death Madella – Ward Linkage*



The ward linkage shows some similar results regarding Nelson Mandela and Religious talks but fails to clearly identify the other major topics.

VIII. K Means vs. Ward Linkage

As seen by the results, K Means provides more meaningful results than Ward Linkage hence we'll use K Means results for further analysis.

IX. C# Visualization

C# visualization of results for the desktop application is shown as. The code uses the Matlab generated result files.

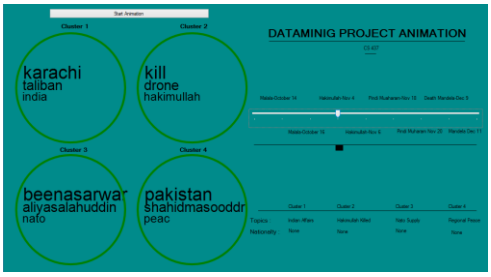


Figure 3 C# Result Visulization

X. Cluster Analysis (K Means)

Using the K Means results we summarize the general discussion topics during the occurrence of events as

Malala - Nobel Peace Prize

14^{th} – 15^{th} October 2013		16^{th} – 18^{th} October 2013	
Cluster 1	Police power in Baloch.	Cluster 1	Terrorism & Peace
Cluster 2	General politics	Cluster 2	Malala Nobel Prize
Cluster 3	General politics	Cluster 3	General Discussion
Cluster 4	Eid Holidays	Cluster 4	Pakistan Army

Topics Discussed: Malala, Nobel Peace Prize, Eid, Pak Army, Terrorism

Hakimullah killed in Drone

4^{h} – 5^{h} November 2013		6^{th} –8^{th} November 2013	
Cluster 1	Terrorism & Karachi	Cluster 1	Politics – Karachi
Cluster 2	Hakimullah Kill (more)	Cluster 2	General Politics
Cluster 3	General Political Talk	Cluster 3	Hakimullah Kill (less)
Cluster 4	General Politics	Cluster 4	General Political Talk

Topics Discussed: Hakimullah killing, Terrorism, Politics

Pindi attach - Muharram

18^{th} – 19^{th} November 2013		20th – 22nd November 2013	
Cluster 1	Beirut Bombing (less)	Cluster 1	General Politics
Cluster 2	Enforce law & security	Cluster 2	Abdallah abduction
Cluster 3	General Talk	Cluster 3	Beirut Bombing (more)
Cluster 4	Rawalpindi Killings	Cluster 4	Rawalpindi Killings

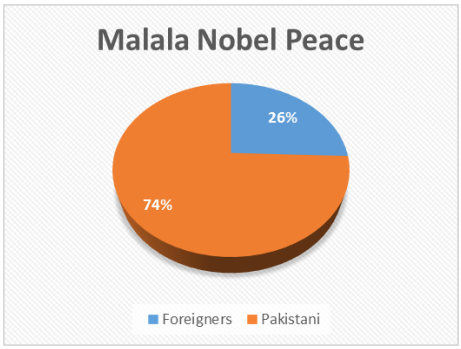
Topics Discussed: Beirut Bombing, Rawalpindi Killing, Amina Abduction, Sunni Shia Conflicts, Law and Security.

Death Mandella

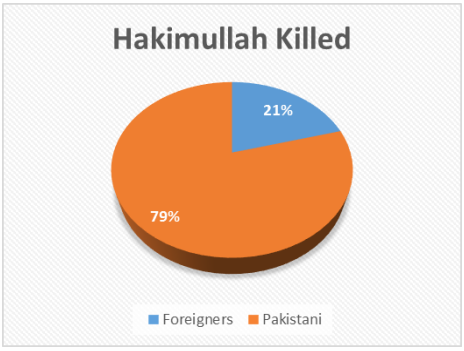
9^{th} – 10^{th} December 2013		11^{th} – 13^{th} December 2013	
Cluster 1	Benchmarking IMF	Cluster 1	General Politics
Cluster 2	Sindh Politics	Cluster 2	Mandella Death
Cluster 3	Pakistan US Relations	Cluster 3	Missing Persons - SC
Cluster 4	Religion talks - Ahmadi	Cluster 4	Pakistan US Relations

Topics Discussed: Ahmadi talks, Mandela Death, Pak-US Relations, and Missing Persons.

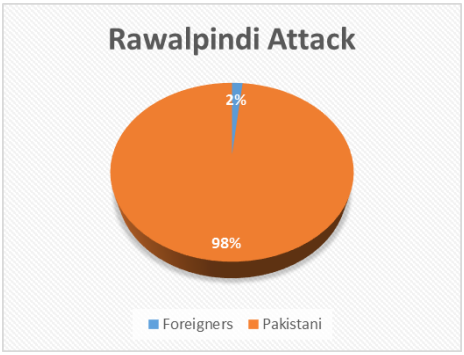
Considering the major events, the amount of foreigner and local population discussing the topic is summarized by the given pie charts.



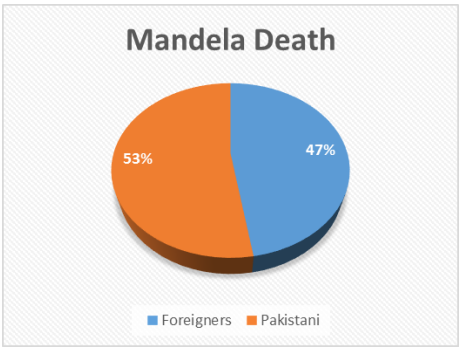
As expected the local journalists discussed the Malala event more frequently than the foreigners.



The results are somewhat similar to the Malala Nobel Peace Prize discussion with a decent amount of foreigners also discussing the topic.



Somewhat strange is the fact that the brutal rawalpindi killings generated less to none interest or attention from the foreign journalists, considering their involvement in other terrorist related news in general.



Mandela death being an international event was discussed equally by both sides.

XI. Conclusions

In this article, we have shown how social media can be utilized to analyze current popular discussions during the duration of an event. Specifically, using the rate of chatter from the popular site Twitter, we constructed a model for finding the importance of an event to foreign and the local population. Also the K Means clustering seemed to produce better results than the ward linkage technique used in Matlab TMG. The stemming algorithm showed its limitations with the local Urdu words and the prefixing technique used, which can be improved in future work.

Acknowledgment

We would like to acknowledge Dr. Fawad Hussain for his guidance, help and contribution towards accomplishment of this report.

References

[1] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1), Jan 2009.

[2] B. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 2009.

[3] W. Zhang and S. Skiena. Improving movie gross prediction through news analysis. *In Web Intelligence*, pages 301304, 2009.

[4] Akshay Java, Xiaodan Song, Tim Finin and Belle Tseng. Why we twitter: understanding microblogging usage and communities. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, 2007.

[5] Ramesh Sharda and Dursun Delen. Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, vol 30, pp 243–254, 2006.

[6] Daniel Gruhl, R. Guha, Ravi Kumar, Jasmine Novak and Andrew Tomkins. The predictive power of online chatter. *SIGKDD Conference on Knowledge Discovery and Data Mining*, 2005.

[7] Mahesh Joshi, Dipanjan Das, Kevin Gimpel and Noah A. Smith. Movie Reviews and Revenues: An Experiment in Text Regression *NAACL-HLT*, 2010.