1A3-IOS-3a-6

Using Social Networks in Political Elections

Ahmed Sameh[*1]

[*1] *Prince Sultan University*

We propose a new real-time analytic tool for elections. The proposed tool collects live tweet streams on frequent time intervals from followers of presidential candidates and computes a mix of qualitative and quantitative indicators to measure our three parameters: Opinion, Influence and Trust spread (metrics dashboard). The tool provides three levels of analysis (algorithms): Surface, Shallow, and Deep depending on analyst wish and the traffic rate of input streams for the purpose of providing real-time mining service. English, French and Arabic tweets are analyzed taking into consideration important context background information.

1. Introduction

One out of every five people worldwide will use a social network this year; one in every four will do so in 2014. Defining users as those who visit a social network at least once a month, 63.2% of the world's internet audience will use social networks. This number is projected to rise to 67.6% in 2013 and then to 70.7% in 2014. Most users have multiple accounts on popular Social Networks and spend fair amount of time reading and writing posts on these networks. Social Networks Analytic tools are used to monitor and analyze posts and activities on these Networks. Social Networks are the new media that people use for exchanging ideas and opinions about many areas such as arts, politics, religion, science, social, and economic issues. In this paper we are drawing our attention to the area of politics. In particular we are analyzing political posts and studying the three 2012 presidential elections: the American, the French, and the Egyptian over Twitter posts as case studies. We collect live tweet streams on frequent time intervals from followers of several presidential candidates and compute a mix of qualitative and quantitative indicators to measure out three parameters: Opinion, Influence and Trust spread. Opinion of followers about the candidate and his/her stand on issues and how this spreads. Influence of the candidate on his followers and how this spreads. Trust measurement among voters and candidates is also of our interest. We provide three levels of analysis: Surface, Shallow, and Deep depending on analyst wish and the traffic rate of input streams for the purpose of providing real-time mining service. English, French and Arabic tweets are analyzed taking into consideration important context background information. We focus more on "Arabic", and present three "Tweet Coloring" algorithms that make use of previously developed NLP tools and graph algorithms such as "Arabic WordNet", "Q-WordNet" ontology for sentiments, "Arabic Lexicons", "Arabic Tweet Corpus", "Max Flow Minimum Cut" in order to speculate voters' inclination and impression about the candidates and the influence of the candidates on the voters. Similar tools and algorithms are used for English and French tweets. On the other hand, three "Edge Coloring" algorithms are used to speculate on opinion, influence, and trust spread (cascading) through Twitter social network graphs. Each one of the three Edge coloring algorithms has a "binary" version (+ve/-ve, Yes/No) and a "continuous" version. The algorithms compute inferred value ratings through polling neighbors and weighted averaging.

2- Tweet Coloring and Text Mining

Text mining is based on "Natural Language Processing- NLP" and "Information Retrieval-IR". On the other hand, an opinion (or regular opinion) is simply a positive or negative sentiment, view, attitude, emotion, or appraisal about an entity or aspect of the entity. Opinion orientation is positive, negative or neutral. The goal of an opinion is either speculative, persuasive, impression, and/or inclination. Opinions matter a great deal in politics. In this work we focus on understanding what voters are thinking, election candidates' attacks, support, etc. Tweet coloring expands on the idea of opinion orientation, and borrows from Web page ranking. We apply similar ideas using text mining algorithms to color tweets according to the degrees of voting or non-voting for the candidate. We collect live tweet streams on frequent time intervals from followers of presidential candidates and apply these tweet coloring algorithms to color tweets to measure opinion support or non-support for a candidate. We provide three levels of analysis: Surface, Shallow, and Deep depending on analyst wish and the traffic rate of input streams for the purpose of providing real-time mining service. English, French and Arabic tweets are analyzed taking into consideration important context background information (e.g. culture of the country, political background, nature of the election process, etc.). We focus more on "Arabic", and present three "Tweet Coloring" algorithms that make use of previously developed NLP tools and graph algorithms such as "Arabic Wordnet", "Q-WordNet" ontology for sentiments, "Arabic Lexicons", "Arabic Tweet Corpus", in order to speculate voters' inclination and impression about the candidate. Similar tools and algorithms are used for English and French tweets.

At the surface level of analysis we apply word-level text mining algorithm called "Word Bag". In this algorithm we have training and testing phases. In the training phase we analyze a collection of Tweets from the Tweet stream and build two word bags "Positive Words Bag" and a "Negative Words Bag" (sentiment words such as great, excellent, horrible, worst, are used to identify which bag to color the Tweet, etc.). This is kind of training building an "Opinion" corpus. The need for this step is evident since Tweeters' language is completely different from standard everyday life text language. Tweeters use their own "slang" language and their own

Contact: Ahmed Sameh, Prince Sultan University, P.O.Box 33866 Riyadh, KSA- asameh@cis.psu.edu.sa

special words and symbols. That is why it is important to build a special corpus for this Tweets language, we called it “opinion” corpus. In the first step of the coloring algorithm, Tweet clustering is used to reduce the dimensionality and categorization of the training tweets according to the political domain. Unsupervised machine learning is used at this step. Tweet pre-processing is the next step, where text cleanup is performed (un-important stop words are deleted), followed by “Tokenization” where roots (stems) of words are produced, followed by POS (Part of Speech Tagging) where syntactic tagging is applied, and finally “word sense disambiguation” is applied. In the training phase a teacher (supervised training) is needed to classify the produced words into one of the two bags: positive and negative. This is called “opinion lexicon generation” stage. Then in the testing phase, tweets are colored positive or negative depending on the number and POS of the +ve/-ve words in the tweets identifies from the two bags. This Tweet surface coloring algorithm has a “binary” version (+ve/-ve, Yes/No) and has no “continuous” version.

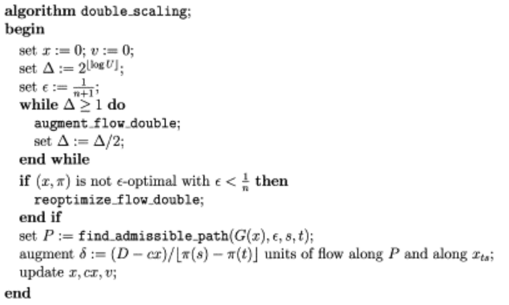
The Shallow level of analysis extends the previous level of analysis by replacing the “bag of words” with a “feature vector”, where the words produced from the training tweets are inserted into a feature vector in the vector space. In the politics domain we choose the best features that best characterize the domain. Then we use a classifier to automatically generate labels from these feature vectors. These labels are thought of as column heads of a database table. This is a tweet level syntactic analysis where phrases along with words are also considered in the feature vectors. This is a continuous version of the surface analysis algorithm above.

The “Deep” level uses the attributes identified in the previous algorithm to create a corresponding database. Now the training set of tweets has been converted into a structured database. At this stage we can apply data mining algorithms such as discovering patterns, answer queries. At this level traditional data and visualization tools are used to analyze and visualize the results. Semantic analysis is considered at this level where ontologies such as “Wordnet” and “Q-Wordnet” are used to understand hidden meanings. Also background context knowledge about the politics domain is consulted during the semantic analysis. Semantic analysis is considered at this algorithm where ontologies such as “Arabic Wordnet” and “Q-Woednet” are used to understand hidden meanings in the Tweets. Arabic Wordnet is classified into nouns, verbs, and adjectives. Once a POST (Parts of Speech) analysis of a Tweet is done, nouns, verbs and adjectives can be identified and used as input into their corresponding Arabic Wordnet ontology and consequently deep understanding of the Tweet becomes possible. Background and context can also be used as in our previous work [1] to help understand the Tweets.

3- Edge Coloring and Data Mining

Twitter followers and following create twitter graphs that can be manipulated using traditional graph theory algorithms to measure the spread (cascading) of opinions, influence and trust in a particular social network graph. We use the following “Edge Coloring” algorithms to discover the spread (cascading) of opinion, influence and trust. They are based on the traditional “Maximum Flow Minimum Cut” family of algorithms. The idea is that we would like to compute the flow of “Opinion”, “Influence”, and “Trust” from one follower to another by analyzing a sequence of Tweet exchange between them. There is considerable literature on graph theory, network optimization, and the minimum cut set problem. The accomplishments of the work reported here are a) to find and implement a practical way of solving large networks for minimum cut sets and b) to discover that the minimum cut set for a large Twitter network was much smaller than what might have been expected given the number of edges in the network. Figure 1 shows the details of the opinion spread algorithm. After using any of the Tweet coloring algorithms above, an edge between two node i and j is either incremented or decremented according to the tweet color ij or ji. At the end of this stage an edge is colored according to the accumulated count over the edge. Large positive corresponds to heavy dark color, large negative corresponds to light blurred color. We use these counts as edge capacities in the well known “Maximum Flow Minimum Cut” problem. To determine where a minimum cut set is to be located, we drew two concentric circles around a central point in the area in question. The source of “Opinion” is assumed to start at an unknown location outside the outer circle, with the intention of reaching an unknown location inside the inner circle. The actual path network considered consists of those edges that have at least one node (i.e., endpoint) between these two concentric circles. The second step in our algorithm is to find a minimum cut set for the graph whose segments have one or both nodes between these boundary lines. This is a well-known problem in graph theory, and it might be expected that there would be many solvers that could be used to obtain a solution to it. However, all but one of the solvers we considered could not find minimum cut sets in networks as large as the Twitter network surrounding Obama. One, the GNET solver, could do so. Several solvers were also Excel-based, and so were restricted by the size limitations in capacity. Others could only accept the problem in the form of a general linear program and their LP-interfaces were unable to handle the problem. As a result, it took significant effort for us to identify just one solver that could find minimum cut sets in networks as large as the Obama’s network. That one was the GNET solver [2]. GNET is designed to handle very large networks, and our experience so far is that it can handle networks of over one million arcs.

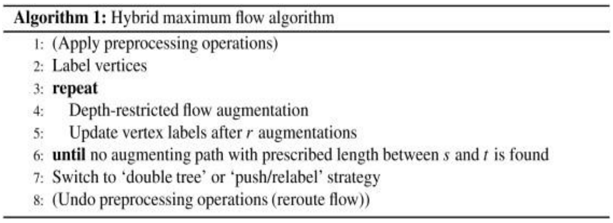
To let the max-flow algorithm find a minimal cut set, the following structure was used. Each of the segments with both nodes inside the ring was given a capacity of counts as described before. An artificial “super-source” node was added outside the outer ring, and an artificial “super-sink” node was added inside the inner ring. The outer node of each segment crossing the outer ring was changed to this super-source node and these segments were given an infinite capacity. Similarly, the inner node of each segment crossing the inner ring was changed to this super-sink node, and these segments were also given an infinite capacity. Finally, an artificial road segment going from the super-sink to the super-source was added, also with an infinite capacity. We then solved this network for its maximum flow and, hence, for a minimum cut set. This minimum cut set gives the smallest number of users that must be there in order to ensure that any "opinion" attempting to penetrate the inner circle, starting on any node from outside of the outer circle, will necessarily encounter the "Candidate" node. It should be noted that there could be more than one minimum cut set, and if there is more than one, the cut sets may or may not include some of the same nodes.



```
algorithm double_scaling;
begin
  set $x := 0$; $v := 0$;
  set $\Delta := 2^{\lfloor \log U \rfloor}$;
  set $\epsilon := \frac{1}{n+1}$;
  while $\Delta \geq 1$ do
    augment_flow_double;
    set $\Delta := \Delta/2$;
  end while
  if $(x, \pi)$ is not $\epsilon$-optimal with $\epsilon < \frac{1}{n}$ then
    reoptimize_flow_double;
  end if
  set $P :=$ find_admissible_path$(G(x), \epsilon, s, t)$;
  augment $\delta := (D - cx)/\lfloor \pi(s) - \pi(t) \rfloor$ units of flow along $P$ and along $x_{ts}$;
  update $x, cx, v$;
end
```

Figure 1: Opinion Spread Algorithm

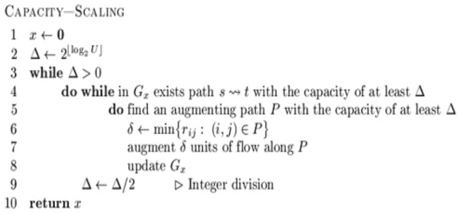
Social Networks tools such as Klout [3] measure one's influence on a network of followers based on his/her ability to drive action within a certain exchange. Kred [4] analyze tweets over the last 1000 days of an exchange. It analyzes what others do because of you. Influence increases when others take action because of one's content. Peer Index [5] measure how active, receptive audience within an exchange. Tweepar [6] computes the ratio of number of followers to the number following.



```
Algorithm 1: Hybrid maximum flow algorithm
(Apply preprocessing operations)
Label vertices
repeat
    Depth-restricted flow augmentation
    Update vertex labels after $r$ augmentations
until no augmenting path with prescribed length between $s$ and $t$ is found
Switch to 'double tree' or 'push/relabel' strategy
(Undo preprocessing operations (reroute flow))
```

Figure 2: Influence Spread Algorithm

Figure 2 shows the influence algorithm that computes a numeric score for a presidential candidate. It measures the volume of waves (cascading of information flow) that his posts generate over the Social Network. The algorithm uses statistical indicators to measure "Influence" of Election Candidates/Voters in Twitter. We need to compute for each Candidate his indicators using the following indices: 1- No of Tweets he sends a day, 2- No of mentions he receives a day, 3- No of Re-tweets his followers make a day, 4- No of people he follows, 5- Eigen-Factor: Weight of "highest-count-of-followers" person who ever Re-Tweeted one of his Tweets, and other factors. Figure 3 shows the "Trust Spread" algorithm. This algorithm has two versions: Binary and Continuous. Its basic structure is: Source polls neighbors for trust value of sink, then the source computes the weighted average of these values to come up with an inferred trust rating; when polled, neighbors return either their direct rating for the sink, or they apply the algorithm themselves to compute a value and return it.



```
CAPACITY—SCALING
$x \leftarrow \mathbf{0}$
$\Delta \leftarrow 2^{\lfloor \log_2 U \rfloor}$
while $\Delta > 0$
        do while in $G_x$ exists path $s \rightsquigarrow t$ with the capacity of at least $\Delta$
                do find an augmenting path $P$ with the capacity of at least $\Delta$
                   $\delta \leftarrow \min\{r_{ij} : (i,j) \in P\}$
                   augment $\delta$ units of flow along $P$
                   update $G_x$
           $\Delta \leftarrow \Delta/2$        ▷ Integer division
return $x$
```

Figure 3: Trust Spread Algorithm

4- Analytic Tool Architecture

Figure 4 shows the system architecture of the proposed analytic tool with metrics dashboard. The "Analytic Engine" communicates with the Communication Module to call the Twitter Server Backend using Twitter APIs to retrieve qualified Tweets for analysis. The retrieved Tweets are passed by the Engine to appropriate analysis module according to the End user needs.

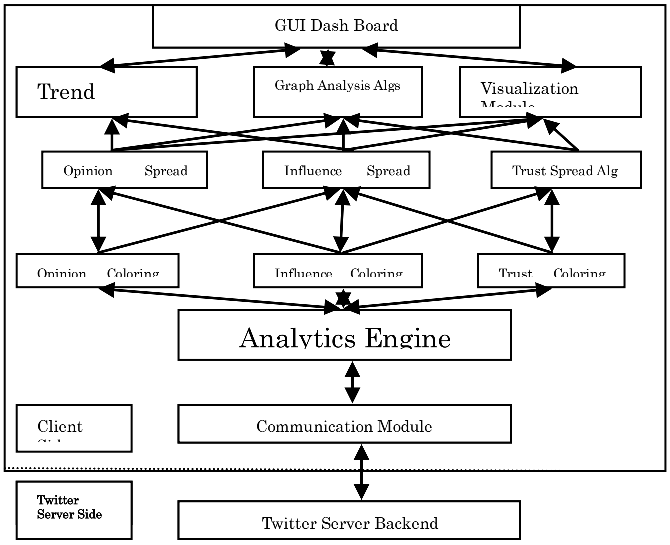


Figure 4: System Architecture

5- Prototype Implementation

A prototype of the analytic tool with metrics dashboard is implemented on-top of "Nodexl" [7]; an open source template for Microsoft Excel that allows automated connection to a social network server and import (Using Twitter APIs) any data stream into the usual Excel environment. Tweet coloring and Edge coloring algorithms are implemented as Excel Macros with selective setting to either surface, shallow, or deep analysis parameters. Influence is measured through some twenty graph related terms such as: Degree centrality, Betweenness centrality, closeness centrality, Bit-ly interestingness, visibility measure, cascade measure, etc. Visualization graphs are provided that allow dynamic filtering, vertex grouping, adjusted appearance (zoom into areas of interest), graph metric calculations, etc. In order to provide real-time mining service even in 'Deep' analysis setup, a parallel cluster farm of duplicate servers is provided in order to help with parallel processing capabilities. Each analytic algorithm will give indication(s) of where the candidate polling numbers are heading, based on its crowdsourced data-set and its analytic algorithm. Each analytic algorithm provides sentiment analysis results with human classification of the sentiments expressed in tweets. Figure 5 shows two screenshots of the implemented prototype, and Figure 6 shows parts of the Macros.

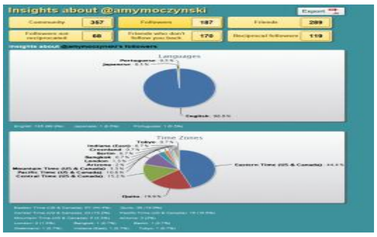


Figure 5: A Screenshot of the Analytic Tool Dashboard

Presidential candidates are heavy weight users of Social Networks. For example, Barak Obama has 14 Million Followers on his Tweeters account, and 10 Million likes on his Facebook account. We have applied our analytic tool on the three elections of 2012- the America, the French, and the Egyptian. We have implemented the above Tweet and Edge coloring algorithms are macros in our tool. Figure 6 shows part of the Tweeter and Edge coloring Macros.

```
Public Sub Shallow-Tweet-Coloring()

    Dim ie As Object
    Dim WebPage As String
    Dim wb As Workbook
    Dim ws As Worksheet
    Dim vWord As String
    Dim current As Workbook
    Dim rng As Range
    Dim lrow As Long
    Dim Rw_Start As Long
    Dim  Colored_Text
    Dim Transl_Text
    Dim MyText As String
    Application.ScreenUpdating = False
    Set current = ActiveWorkbook
    lrow = current.Sheets(2).UsedRange.Rows.Count
    Shallow-Tweet-Coloring_Text =
Application.InputBox("Enter the Row no. for
Coloring On 'A' for all.", "Initialization")

        Set Coloring_Text = Nothing
    End If    '--------------- For Shellow("fa" in
the url = Shallow, and if using Shallow
Coloring it will be "iw") --------------------
   For i = Rw_Start To lrow
    MyText = ""    Set ie =
CreateObject("InternetExplorer.Application"
)
    'vWord =
ActiveWorkbook.Sheets(2).Range("AM" &
i).Value
    WebPage = "http://Shallow-Coloring
com/#fa|ar|" &
ActiveWorkbook.Sheets(2).Range("AM" &
i).Value
    ie.Visible = False
    ie.Navigate WebPage
    Application.Wait (Now +
TimeValue("0:00:4"))
    Do Until ie.ReadyState = 4 'For STATE as
```

```
Public Sub Surface-Color()
   Dim ie As Object
    Dim WebPage As String
    Dim wb As Workbook
    Dim ws As Worksheet
    Dim vWord As String
    Dim current As Workbook
    Dim rng As Range
    Dim lrow As Long
   Sub ProcessColumn()
  Do Until ActiveCell.Value = ""
  If ActiveCell.Value > 10 Then
   Selection.Interior.Pattern = xlCrissCross
  Else
   Selection.Interior.Pattern = xl
 End If
   Selection..Surface-color.Offset(1,
0).Range("A1").Selec
Loop
End Sub
Sub ListWorkbooks()
Dim wb As Workbook

    ActiveWorkbook.Sheets(2).Range("AB" &
i).Value = MyText
    Set Surface-Color_Text = Nothing
    Close_IE
    Set ie = Nothing
    Next i
End Sub
 ' Close_IE to close all instances of IE.
 Function Close_IE()
    Dim objWMI As Object, objProcess As
Object, objProcesses As Object
    Set objWMI = GetObject("winmgmts://.")
    Set objProcesses = objWMI.ExecQuery( _
        "SELECT * FROM Win32_Process
WHERE Name = 'iexplore.exe'")
    For Each objProcess In objProcesses
        Call objProcess.Terminate
    Next
    Set objProcesses = Nothing: Set objWMI =
Nothing
End Function
ie.document.getElementById("result_box").inn
```

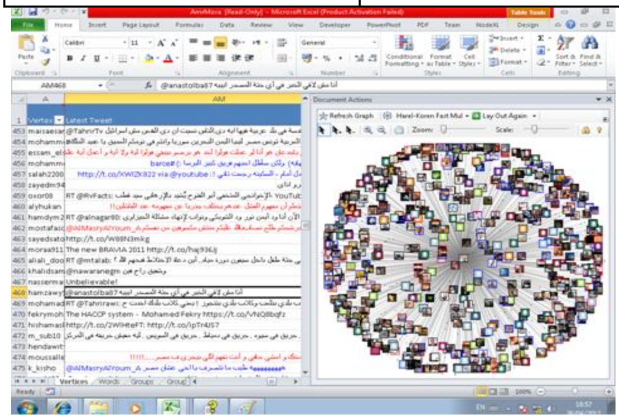


Figure 7: The Egyptian Election Case Study

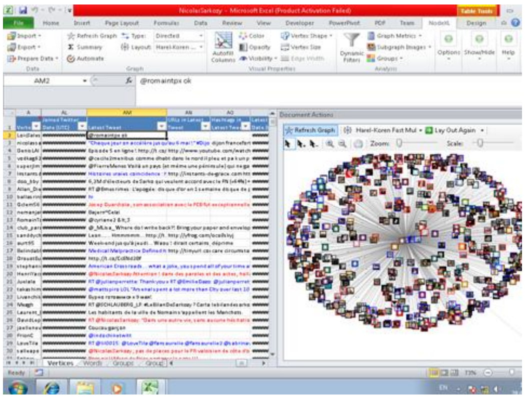


Figure 8: The American Election Case Study

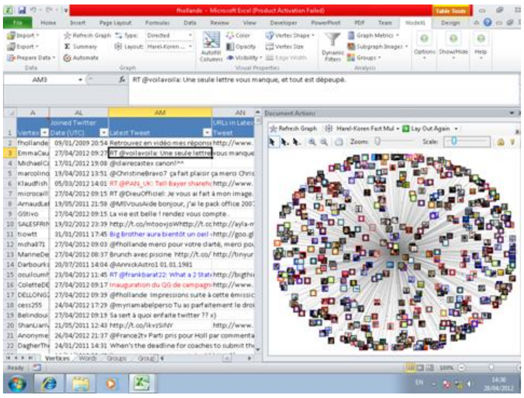


Figure 9: The French Election Case Study

6- Conclusion

The proposed analytic tool can serve as a Gallop for any "Ad Hoc" queries by incorporating relevant color-related algorithms to color code the answers in the Tweets and the edges in the graphs. It is also applicable to other domains of applications such as "Stock Market Predication", "Public Opinion", "Customer Voice", "Service Benchmarking", and "Blog Analysis". Our future work will add translation-to-Arabic capability from any language to Arabic, archive posts over a longer period of time for better history (finger print), and will tackle other social networks such as Facebook, Flickr, Youtube, etc.

References

[1] Sameh A., "A New Twitter Client for Participating-in and Mining Social Networks of Other Languages/Cultures", to appear in the Proceedings of the IDA 2013, UK October 17-19, 2013

[2] Kodama Y., Kudoh T., Takano R., "GNET gigbit Ethernet Network Testbed", Proceedings of the 2004 IEEE International Conference on Cluster Computing. 2004

[3] www.Klout.com/

[4] www.Kred.com/

[5] www.peerindex.com

[6] www.tweepar.com

[7] www.nodexl.codeplex.com