International Journal of Business Quantitative Economics and Applied Management Research
Volume-6, Issue-9, 2020 ISSN No: 2349-5677

ANALYZING AND VISUALIZING TEXT SIMILARITY COMPUTATION MODEL FOR IDENTIFYING BEST TIMES TO TWEET BASED ON K-SHELL ALGORITHM

Dr.Imad Ismail Nazzal
Technology and Applied Sciences Department
Al-Quds Open University
Palestine

Abstract

Analysis Twitter information are becoming an important section of most companies, most companies usually looking for best time of day to post a tweet to get the most customer visibility and responses, companies usually used real time analysis tools to follow up the customer's reaction or responses. In this article, the author provides a new approach of real-time analytics to find the best time to publish a tweet by studying the most time influential between different node level. The writer used a random set of tweets for Al-Jazeera channel (Ajarabic). The author implemented and developed API tools to collect a random set of data contains tweet time, tweeters user, re-tweeters users, and their followers up to three node level edge. The K-Shell Decomposition algorithm was used to calculate a twitter edge weighted. These findings were represented by using Graph visualization tool (Gephi). Our high-level goal is to investigate how time stamp is different between each re-tweet in different node edges. We designed the prototypes to calculate a twitter edge weighted between different node level. Moreover, this paper aimed to investigate the best time to publish tweet and proposed the most accurate follower of tweets.

Keywords: Social networks, twitter, re-tweet, API, Graph visualization tool, Java.

I. INTRODUCTION

Twitter is a social networking tool that offers a micro blogging service, which allows users to post Tweets of user's status with a maximum of 280 characters per message (Naseemet al.,2020). Nevertheless, these tweets are then posted online and can be downloaded the public sees it. Micro blogging has become a common connection among Internet users. Twitter users can send their Tweets, follow other users' Tweets or participate in a broader discussion on a topic or event, Twitter is generally referred to as a "micro blogging" tool (Luo et al.,2018). In the literature of (Hellsten et al., 2019) explained that in last decades, visualization and graphs spread rapidly through social media, many studied has explained the importance the importance of visualization and graph analysis in real time for generating a recommendation at Twitter. In other hand, visualization analysis in real time for retweet helping to contribute a new approach to many companies to augment the connection between followers and content (Sharma et al.,.2016). Companies used different commercial tools and system that analyze retweet data to

suggest the most appropriate time to publish retweet based on follower's interaction (Goel et al.,2016). Companies used a Prime tool, the tweriod site, and the Follower Wonk tool on Twitter to analysis the appropriate time to publish their retweet (Lukasik et al.,2016). Those tools only analysis the first level of retweet a node time step without giving an importance to second level or other node between followers. In this paper, we investigate to calculate a twitter edge weighted between different node inherent to a re-tweet between follower based on timestamp to identify the sets of datasets for classifying the importance of a node edge to investigate the best time to publish tweet and proposed the most accurate follower of tweets.

II. RESEARCH STATEMENT

Organizations and companies have many challenges to find an appropriate time to post their retweet (Jianqiang et al.,2018). The most analysis tools only analysis the first level of retweet a tweet, in other words, when follower retweet the tweet other follower might be retweeting the same tweet without following the main source. Analysis the first node does not always give satisfactorily result (Murthy et al.,2019). Considered the second or third level of retweet users giving more accurate data result. Different node edge in twitter may influence more than the first edge tweet. allow the companies to have the accurate decision based on a properly analysis to analysis the timestamp for their follower and follower who follow their follower. In this research we designed the prototypes based on K-Shell algorithm to analysis up to three node Level edge to find appropriate time to publish a tweet by studying the most time influential between a tweeting and re-tweeter user for a random set of tweets up to three nodes.

III. RELATED STUDIES

Twitter has a restriction that prevent companies to post more than 140 characters each tweet, May companies has massive followers and posting many tweets on their twitter account. Therefore, many companies post similar multiple postings tweet every few hours without considers the importance of time stamp of their retweet (Ranjan et al.,2018). The study of (Fayoumi et al.,2017) suggested that the appropriate time to post retweet in twitter is from 7 am to 8 am, and from 5 to 6 pm. The worst times are Monday and Friday. Moreover, the study of (Hasan et al.,2018), found that twitter timeline analysis has many challenges to find accurate the time stamp especially when tweet has many followers, Maynard found that the visualization is suitably technology in monitoring and following the twitter stamp time. According to (Pratap et al.,2018), showed that graph real time analysis is powerful tools to measure how information popularize by developed algorithm to present data in graph clustering the real time data in different nodes to overcome the massive data and information. In other hand the (Tyagi et al., 2019), Implemented an experimental study that summarize the overall retweet in different node level. The authors developed a "wisdom of the crowds" algorithm that measures of central tendency (node) and observation the retweet chart and generate real time prediction system to detected the time stamp between retweet and predicted most influence time that follower retweet other tweet and generate dynamic timestamp between different node and different retweet. While the study of (Lukasik et al.,2016), investigated that twitter real time analysis helps

companies to constant and dynamic road to prophesy the spread of their tweets. The author's designed visualization tool consists on Twitter Trend Momentum approach that index tweet into different nodes depending on timestamp to detected the strength sentiment tweet. The study showed that visualization has become a complementary ingredient in time analysis. (Belmonte et al.,2014), proposed a Graph visualization technique shows a that characterize the strength the time stamp of any re-tweet that being debated on twitter. The authors proposed method of real-time obligation by developed a visualization technique in creating a real-time classification on Twitter for the preferable time stamped that can publish the tweet there is insufficiency of algorithms technique and approaches for deep and fast analysis of social media in real-time. The study of (Jain et al.,2018), explained that many organizations put more effort and time on social media sites such as twitter. In other hand, the massive data and follower inspire organization to used analysis tools to acquire a useful knowledge in real time. The author used a Clustering algorithm to create two different cluster for each retweet to detect the weights of time stamp with index the tweets posted by follower to keep track of time stamp for each retweet and detected the highlight topic between follower with time stamp.

IV. **METHODOLOGY**:

In this paper, we implemented a Java application based on library for the Twitter API in order to fetch and download the tweets using Twitter API based on Twitter 4J v 4.0.5. Twitter API has restriction to download data and request data, the twitter API allows sending 350 query every hour.

4.1 PAPER DESIGN

Our prototype integrated to Twitter API and it able to send 350 queries every hour. The tools perform queries for 41 days to collect and extract retweet ID, re-tweet followers from three node edge with their retweet and recorded each re-tweet with a unique stamp-time. Each query result sent to the serialization java file that store re-tweets data in two datasets in MySQL database automatically. Moreover, the tools have been running for 41 days and collected all the re-retweet and timestamp for each retweet up to three nodes, A Gephi visualization tool v 0.9.1 based on Net Beans platform has been proposed to create an accurate visualization paragraph that describes the appropriate time and the significant time of retweets that being discussed Recommended font sizes are shown in Table 1.

4.2 DATA COLLECTION

The data were collected from Al-Jazeera channel (Ajarabic). We collated 11,000 tweet that was retweet from 851,15 followers. All data were collected from 10 March 2020 till 20 April 2020. The tweets were collected were automatically store in two datasets based on K-Shell Decomposition algorithm; we labelled the tweets in order to establish the relationship between retweet with timestamp for each re-tweet ID and followers ID up to three nodes. Due to the difference tweet time and for accuracy timestamp, we used Greenwich time zone (GMT time).

Consequently, our two datasets consisted of 11,000 tweet that was retweet from 851,15 followers. Previous research has focusing to fetch tweets and real time analysis depends on first node level, in our research we collected data up to three node level for each tweet. After fetching the desired data and stored in two datasets, the table (1) shows the data that we were able to download.

Dataset	Users	Followers	Re-tweeted
Ajarabic	21,201	851,59	11,000

Table (1): Sample data

4.3 IMPLEMENTATION

1. **Analysis Process**

In this research, we implemented and used K-Shell Decomposition algorithm, in order to analyze our datasets into different three k-shell levels based on timestamp between each retweet and follower. The figure (1), shows the frequent node between each retweet up to three level.

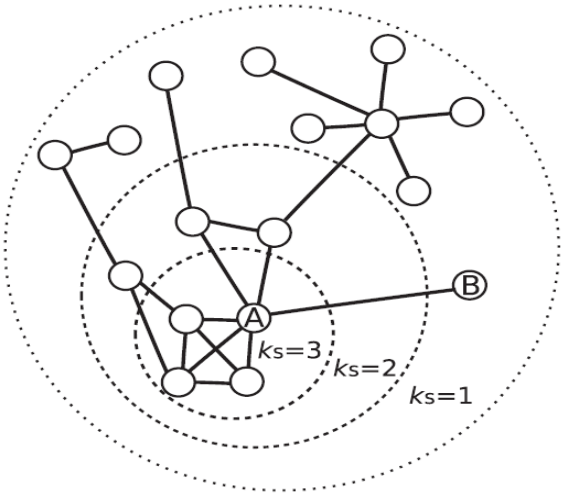

Figure 1: K-Shell algorithm

Classified the frequent primary tweet ID, follower ID and nodes with degree k = 1, and we marked the value of re-tweet the same retweet between the follower as k (i+1) = to 3. The collected date marked in two datasets where the node from one level to three level, in other words where K degree less than 3. After classified all frequent re-tweeter users with k=1, we proceed to the next step to classify the all nodes with degree k = 2, and mark the value as , and

so on up to remove classify re-tweeter users from the network all nodes with degree k = 3, and mark the value as

The author proposed the weighted of degree of a node based on k-shell decomposition $\sum_i^{ki} v$. In particular, the first weighted degree mark as (i=1). Where Ki is a degree of node for each tweet ID and followers ID. While, Wij represented the value of timestamp for each retweet based on primary tweet and the other node level. For this purpose, we calculated the weighted for each node by applying this Ki= $k^i(\sum_i^{ki} w$.

2. Analysis Process Method

The data represented and analysis by visualization graphs where K is a collection of retweet vertices and W is a collection of node connection between followers and tweet based on timestamp up to three nodes. Each vertex was stored in table in MySql as main node. All the data was stored in two datasets and classify by using K-shells to different cluster. We assigned the data to the (wi ,wj), where we can calculated and weighted the radius between tweets nodes and visualization into different cluster by classified that data based on timestamp. The algorithm represents the data to different layers and cluster up to three node and displayed the data in different circular shells to demonstrate the set of vertices for each retweet between the followers based on highest edge in whole the network and weight the appropriate time.

The research implemented different rules to calculate the weight of each tweet by using the formula 2n2, where "n" symbolizes the shell edge, our tools used to store data in serialization java file then transferred and stored the data in MySQL. The researcher used to mark each tweet to avoid the duplication. Thus, retweeting configuration of retweet follows an ascending order.

Tweet weighted by using K-Shell algorithm

K←0
Ki←i
While | Ki |←0 do
While there exist a node Wj in ki with degree ≤k do
Delete v and its edges
End while
Ki+1← Wij
k← k+1
End while

4.4 VISUALIZATION

i. Analysis Activity at First Level node

According to figure (2) the result showed that the followers at first level node retweet showed 70% of re-tweeter user were very active at 21:30 and the low active re-tweeter user were at 12:23 pm. Which it means the best of publish tweet is 21:30 where (70%) of the follower were able to see the tweet and re-tweet it. Moreover, the time at 12:23 pm less (10%) of the follower who was active.

International Journal of Business Quantitative Economics and Applied Management Research
Volume-6, Issue-9, 2020 ISSN No: 2349-5677

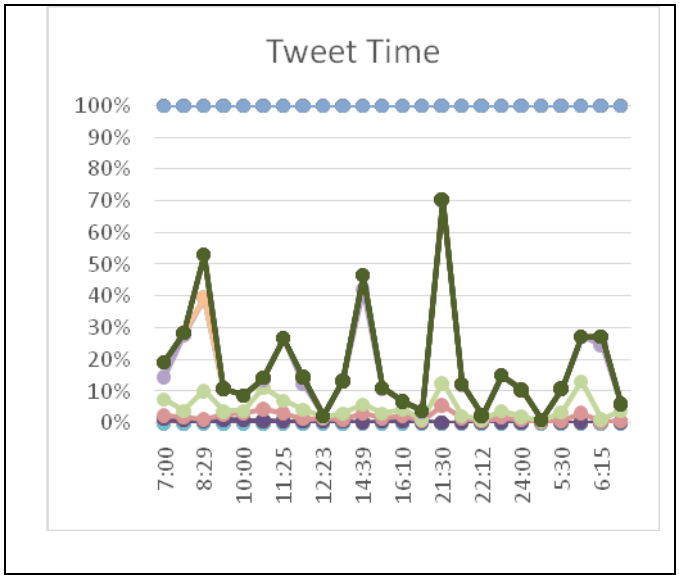

Figure 2. Real time re-tweet Activity first level node.

ii. Analysis Activity all different level node.

The result showed in figure (2), that followers are very active between 16:10 to 21:00, where the follower are less active in the morning. The paper proposed that appropriate time to publish the tweet should be around 16:00-21:12. Moreover, the study suggested the significant time to post the tweet should be after 16:00 pm.

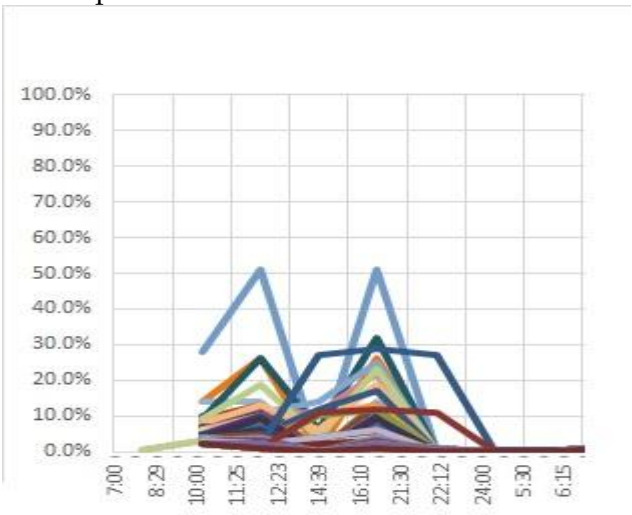

Figure 3. Real time re-tweet Activity all different level node.

According to the figure 3, showed at 10:00 am the users were around 30 % very active, and 14:39 to 21:30 were the most time the user's active, meanwhile, the users at 5:30 am to 7:00 am were

less active. Which it means the channel should not publish at those time if they want their tweet reach more users.

iii. Representing the weight of edges in graph.

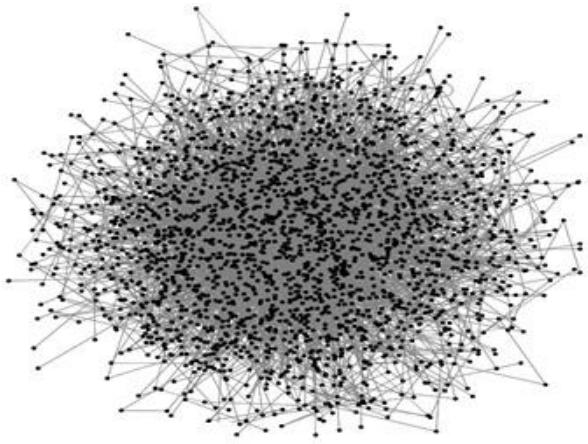

Figure 4. Distribution node tree.

According to Figure 4, the figure showed tweet messages researched different users between them in different level layer among whole networks.

iv. Calculate the weight edges tweet among follower user.

The figure 5, showed the sample of weight of tweet number (715) among follower users, the green line showed the specific tweet with a re-tweeter user in each level node up to three level.

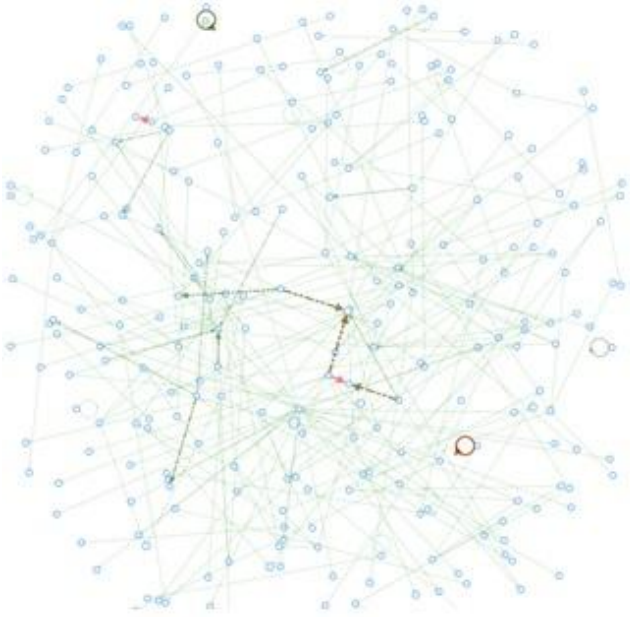

Figure 5. Weight of edges represents Tweet number (715)

v. Re-Tweet a Tweet spread first level among tweet number (715)

The figure (6), shows the path for each tweet in whole network, the Red line in the figure shows how the message spread in different level between followers, and display how follower retweet the tweet in different level from different users. While, the green line in the figure display the follower user in second or third node level retweet the tweet. The figure illustrates how the message interacts with followers at different times in different level. Consequently, the study showed that channel should give more attention to study other users' activities, who are not in the first node and primary follower

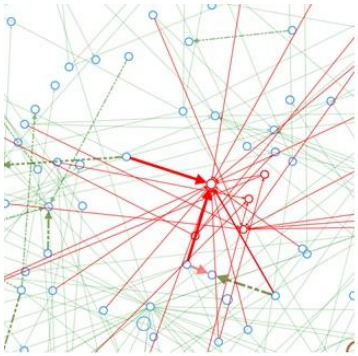

Figure 6. The Node edge among re-tweeters.

V. CONCLUSION

In this research, we developed Java application prototype to analyses the interacted between followers for three level nodes to find the appropriate time to publish the tweet. While the companies and organization are focusing to analysis retweet timestamp based on first node level. The prototype allows to retrieve and analysis data from different node edge. 11,000 Re-tweeted randomly selected and stored in two different datasets. Visualization graph was used to analysis and represented data. The tool allows the organization to choose the appropriate time to publish tweet. The study indicated that a difference of analysis the first node and analysis of other node level. The study showed 21:30 was the significant time to publish tweet in first level. In other hand, the study showed that the time between 16:00-21:12 is the best time to tweet. In the future, we will improve our tool by adding more variable such as follower location, follower behavior. Moreover, increase the analysis efficiency by providing sentiment analysis.

REFERENCES

1. Belmonte, N. G. (2014, October). Extracting and visualizing insights from real-time conversations around public presentations. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 225-226). IEEE.
2. Fayoumi, A., Jackson, C., Lewis, C., Straw, J., Sharpe, J., & Nicol, D. (2017). What they are Tweeting about me?: social media data analytics with geographical visualisation.. Burnap, P., Rana, O. F., Avis, N., Williams, M., Housley, W., Edwards, A., ... & Sloan, L.

(2015). Detecting tension in online communities with computational Twitter analysis. *Technological Forecasting and Social Change*, *95*, 96-108.

3. Goel, A., Gautam, J., & Kumar, S. (2016, October). Real time sentiment analysis of tweets using Naive Bayes. In *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)* (pp. 257-261). IEEE.
4. Hasan, M., Orgun, M. A., &Schwitter, R. (2018). A survey on real-time event detection from the twitter data stream. Journal of Information Science, 44(4), 443-463.
5. Hellsten, I., &Leydesdorff, L. (2019). Automated analysis of actor–topic networks on twitter: New approaches to the analysis of socio-semantic networks. Journal of the Association for Information Science and Technology.
6. Jain, A., Gupta, A., Sharma, N., Joshi, S., & Yadav, D. (2018). Mining Application on Analyzing Users' Interests from Twitter. In Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT)(pp. 26-27).
7. Jayaram, M., Adams, C. E., Friedel, J. S., McClenaghan, E., Montgomery, A. A., Välimäki, M., ... & Zhao, S. (2019). Day of the week to tweet: a randomised controlled trial. BMJ open, 9(4), e025380.
8. Jianqiang, Z., Xiaolin, G., &Xuejun, Z. (2018). Deep convolution neural networks for twitter sentiment analysis. IEEE Access, 6, 23253-23260. Hellsten, I., &Leydesdorff, L. (2019). Automated analysis of actor–topic networks on twitter: New approaches to the analysis of socio-semantic networks. Journal of the Association for Information Science and Technology.
9. Kumar, P., & Sinha, A. (2016, October). Real-time analysis and visualization of online social media dynamics. In Next Generation Computing Technologies (NGCT), 2016 2nd International Conference on (pp. 362-367). IEEE.
10. Lukasik, M., Srijith, P. K., Vu, D., Bontcheva, K., Zubiaga, A., & Cohn, T. (2016). Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Vol. 2, pp. 393-398).
11. Luo, Z., & Liu, X. (2018). Real-time Scholarly Retweeting Prediction System. In Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations (pp. 25-29).
12. Maynard, D., Roberts, I., Greenwood, M. A., Rout, D., &Bontcheva, K. (2017). A framework for real-time semantic social media analysis. Journal of Web Semantics, 44, 75-88.
13. Murthy, J. S., Siddesh, G. M., & Srinivasa, K. G. (2019). A Real-Time Twitter Trend Analysis and Visualization Framework. *International Journal on Semantic Web and Information Systems (IJSWIS)*, *15*(2), 1-21.
14. Pratap, A. R., Prasad, J. V. D., Kumar, K. P., & Babu, S. (2018). An investigation on optimizing traffic flow based on Twitter Data Analysis. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 320-325).

15. Ranjan, S., Sood, S., & Verma, V. (2018). Twitter Sentiment Analysis of Real-time Customer Experience Feedback for Predicting Growth of Indian Telecom Companies. In 2018 4th International Conference on Computing Sciences (ICCS) (pp. 166-174).
16. Sharma, A., Jiang, J., Bommannavar, P., Larson, B., & Lin, J. (2016). GraphJet: real-time content recommendations at twitter. *Proceedings of the VLDB Endowment*, *9*(13), 1281-1292.
17. Tyagi, P., &Tripathi, R. C. (2019). A Review towards the Sentiment Analysis Techniques for the Analysis of Twitter Data. Available at SSRN 3368718..
18. Zarco, C., Santos, E., &Cordón, O. (2019). Advanced visualization of Twitter data for its analysis as a communication channel in traditional companies. Progress in Artificial Intelligence, 1-17.

ACKNOWLEDGMENT

Analysis Twitter information are becoming an important section of most companies, most companies usually looking for best time of day to post a tweet to get the most customer visibility and responses, companies usually used real time analysis tools to follow up the customer's reaction or responses. we will improve our tool by adding more variable such as follower location, follower behavior. Moreover, increase the analysis efficiency by providing sentiment analysis.