A Methodology In Processing Descriptive Analytics Using MMDA Traffic Update Tweets, Tokenization And Classification Tree In Discovering Knowledge

Tristan Jay P. Calaguas, Menchita F. Dumlao

Abstract: Traffic on National Capital Region of the Philippines is going as one of many problems facing by the local government and Filipino citizen who are residing in Metro Manila. In addition, a Filipino citizen that is working in Metro Manila is experiencing a waste of Twenty – Eight Thousand hours in traffic which results unproductivity. Due to traffic that causes long commutes it take away an individual from exercise activities that results fatigue in their health. In relation with this, due to lack of exercise that causing by the traffic, each year, One Hundred Seventy Thousand Filipinos die from cardiovascular diseases up from Eighty Five Thousand more than Twenty years ago, according to 2009 study by the Department of Health (DOH). Population increase is one of many causes of traffic in Metro Manila. As population is growing, the more car riders and commuters volume will be in the road including delivery trucks, Pedi cabs, jeeps, and provincial buses that signify that there is a high employment rate in the country that causes traffic. However, to sustain the public needs, MMDA is the government agency that provides public services to Filipino citizens through providing updated public traffic information. For past years, MMDA used Telephony lines and Television Broadcasting for traffic information dissemination, which is very costly in maintenance that made them to adopt Twitter to post Traffic updates and advisories to the public .Since, this government agency uses Twitter in disseminating information through posting tweet, there is a need for a methodology on how these tweets will analyze so that citizens will have an insight in decision making to avoid specific time of traffic in metro manila. From this condition, the researcher will adopt the use of MMDA tweets as the primary data source and apply the CRISP as the knowledge discovery standard processes that to be used in building methodology for descriptive analytics. In this experimental research several processes were used to convert the semi structured MMDA tweets into structured data matrix. SQL was used for storing, retrieving and pattern matching, while PHP string functions were used to tokenize the tweet and transform it into array so that the tokens can store in database using iterative structure. After loading all token to its specific table we abled to have a data matrix that comprised of time, routed roads, traffic status and day information that was used in data mining to discover knowledge. Lastly we used J48 Classification Algorithm to classify the time usually the traffic happens in many routed roads from NCR. As the result we discovered that from Eight O'clock to Nine Forty One in the morning the commuters are experiencing a traffic and from One O 'Clock in the afternoon to Eight O'Clock in the evening the commuters are also experiencing a traffic in C5 North Bound to South Bound and Edsa North Bound to South Bound every Tuesday and Friday with the accuracy of 75.72%.

Index Terms: tokenization, classification, tweets, traffic, methodology, knowledge discovery, update traffic

1 INTRODUCTION

Traffic on National Capital Region of the Philippines is going as one of many problems facing by the local government and Filipino citizen who are residing in Metro Manila. In addition, a Filipino citizen that is working in Metro Manila is experiencing a waste of Twenty – Eight Thousand hours in traffic which results unproductivity [1]. Due to traffic that causes long commutes [2] can take away from exercise times that results fatigue to individual [3]. In relation with this, due to lack of exercise that causing by the traffic, each year, One Hundred Seventy Thousand Filipinos die from cardiovascular diseases up from Eighty Five Thousand more than Twenty years ago, according to 2009 study by the Department of Health (DOH) [4].

- *Tristan Jay P. Calaguas is currently pursuing Doctorate degree program in Information Technology in AMA University, Philippines, and currently an Information Technology Faculty Member in The Philippines Womens University. E-mail: calaguas26@yahoo.com*
- *Dr. Menchita F. Dumlao is currently the Information Technology Chair Woman in Information Technology and Research Director in The Philippines Women's University, Philippines*

Population increase is one of many causes of traffic in Metro Manila. As population is growing, the more car riders and commuters volume will be in the road including delivery trucks, Pedi cabs, jeeps, and provincial buses that signify that there is a high employment rate in the country that causes traffic [5]. However, to sustain the public needs, MMDA is the government agency that provides public services to Filipino citizens through providing updated public traffic information. For past years, MMDA used Telephony lines and Television Broadcasting for traffic information dissemination, which is very costly in maintenance that made them to adopt Twitter to post Traffic updates and advisories to the public [6]. Since, this government agency uses Twitter in disseminating information through posting tweet, there is a need for a methodology on how these tweets will analyze so that citizens will have an insight [7] in decision making to avoid specific time of traffic in metro manila. From this condition, the researcher will adopt the use of MMDA tweets as the primary data source and apply the CRISP as the knowledge discovery standard processes that to be used in building methodology for descriptive analytics.

2 RELATED WORKS

Organization will have an insight through dealing with huge amount of information by performing the four tasks of knowledge discovery processes. These are the data collection, data cleaning, data analysis, and data application [8]. There are many ways from traditional methods on how they can collect data for decision making such as surveys, observation, and interviewing but the problem from this traditional method

is that sometimes it is not complete or sufficient to fulfill the overall requirements of an organization compared to computerized data collection method such as the use of modern Short Message Services due to being pervasive [9]. When Live Web was invented and implemented by many Information Technology experts, most websites were become interactive. All web application such as Wiki, Social Networking Site, Online Learning Management, Blogs and podcasting were contributing an efficient way on how this organization to do their activities such as coordination, communication, document processing, knowledge transfer and private document processing [10]. However, to make these technologies is valuable and usable, agencies and industries uses these for their public services. In the Philippines different traffic management groups uses Twitter for their traffic information distribution one of this is the MMDA. Twitter is a micro blogging site where the blogger can post short message that comprises of 140 characters [12]. To make tweets usable many researches were developed such as the sentimental analysis where it analyzes whether the commentator has a negative or positive thoughts and tone in a subject. Using these computer processes, the data analyst can have an insight if the commentator is sad, angry, happy, disappointed, surprised, proud, in love or scared [13].

3 OBJECTIVES

The main objective of this study is to proposed a methodology where the CRISP knowledge discovery standard processes will be applied to acquire data matrix that to be used in data exploration for Descriptive analytics model that uses MMDA tweets as the data source. The following listed specific objectives are needed to achieve.

1. To identify the problems that will encounter in transforming semi structure tweet into structured data matrix in which it can be used for data exploration.
2. To prove that the tokenization process can use in transforming semi structured MMDA tweets into structured data using PHP string function, SQL and Database Management System as well as in data cleaning and data preprocessing
3. To identify the precision, recall and f-measure of each time traffic where commuters can experience every Tuesday and Friday in NCR using J48 classification algorithm.

4 METHODOLOGIES

The researcher adopted the Cross Industry Process Model of Knowledge Discovery standard processes that comprises of Six phases [7]. In this experimental research, only Four out of Six Phases was applied since our intention is just only to simulate the proposed methodology to model the predictive analytics for traffic in National Capital Region in the Philippines. The four phases performed are, the business understanding; data understanding; data preparation; and data modeling.

4.1 Phase 1 – Business Understanding

For Phase 1 of Cross Industry Process in Knowledge Discovery Standard process, the researcher decided to create a model that can describe if the Moderate to Heavy Traffic status in circumferential roads on National capital Region of the Philippines will be experienced at Eight O'clock to Nine Forty One in the morning and One O'clock to Eight O'clock at prime meridian. The researcher selected the Twitter as the primary data source of traffic congestion information due to its timeliness and completeness. The MMDA tweet includes place of traffic, exact time of traffic, date that can be converted to day using spreadsheet software such as Microsoft Excel by changing the date format cells, and traffic status that has "Moderate M", "Moderate to Heavy MH" and "Low L"', values, tweet identity number and cause of traffic that is written from tweet advisory.

4.2 Phase 2 – Data Understanding

In Phase 2, the researcher used R application software for collecting MMDA tweets, data framing and File writing. These tasks will required the researcher to register their account in Twitter App developer site for token identity code that is primarily requirement in accessing MMDA tweets. After the MMDA tweets were collected, framed and enabled to generate spreadsheet file that comprises of tweet, viewer reply, date, uniform resource locator, and tweet identity fields as data matrix, we looked and identified irrelevant attribute, these are, viewer reply and uniform resource locator fields. The researcher only selected the date, id, and MMDA tweets while the other fields were deleted. After deletion of irrelevant fields from data frame, the researcher loaded manually the remained comma separated value format fields to MySQL Database table with the name 'tweet' and has three attributes these are the id, tweet and date with 8,790 record set that were collected from the date of September 22, 2016 to November 5, 2016. The researcher found that there are dissimilar syntaxes from loaded MMDA tweets that are written in tweet field. To solve the issue, the researcher selected tweets from record set using the SQL command *'SELECT*FROM tweet WHERE tweets LIKE '%RT MMDA Traffic Update: As of%'* to extract MMDA tweets with similar syntax. As the result, the researcher abled to extract MMDA tweet with similar syntaxes comprised of time, places and traffic status and they stored these into table with the name of 'traffic status'.

4.3 Phase 3 – Data Preparation

4.3.1 – Trimming of 'RT MMDA: Traffic Update: As of'

In Phase 3 Data Preparation, After the MMDA tweets with similar syntax were loaded into table that has name 'traffic status', we retrieved these tweet again together the optimization of the *ltrim PHP function* to trim the group of tokens from MMDA tweet with similar syntax. The trimmed unnecessary group of tokens from the short message comprises of set S = {RT, MMDA:, Traffic, Update:, As, of}, as the result, the time, places and statuses remained while the unnecessary tokens were deleted.

4.3.2 – Keeping the information intact and converting tweet into array

To keep the time traffic, tweet id intact to places that are written on one tweet, we *used string reverse and string replace PHP string function* to replace open parenthesis special character into traffic time and the closed parenthesis special character symbol into tweet id then we converted these into array separated by comma using *explode PHP string function* and we also applied the for-loop iterative structure to access and migrate all tokens from array into database table with the name 'filters'.As the result, the stored tokens from token field of filters table is comprising of place,

time traffic, traffic status with Three categorical value “Moderate M”, “Moderate to Heavy MH” and “Low L”, and tweet id. All of these values were stored in the same column with the name ‘token’.

4.3.3 – Data Migration

During the place, status, time, and tweet identity were stored in filters table, then we assigned synchronization id field as the primary key at auto increment mode. As the result of test trial, the primary key value of the synchronization id of place value was 1, while status value was 2, time value was 3, date value was 4 and tweet identity numeric value was 5 followed by the same value series such as second place value 5, status of second place value is 6 and so on. After observations, we created Three (3) database tables with the name of “status”, “time” and “date”. From the discovered value series from database filter table, we used SELECT SQL command together the WILDCARD for pattern matching to retrieve and transfer the values in Four (4) cases, these are the following. The time attribute value from filters table into time table with Two (2) attributes that comprises of time identity number and time fields; the traffic status attribute value from filters table into status table with Two (2) attributes that comprises of status identity number and status fields; the date traffic attribute value from filters table into date table with Two (2) attributes that comprises of traffic date identity number and date fields. To address the places’ time traffic, status traffic and day traffic when transformation and loading task was performed, we looked and observed the synchronization identity number of place that is located in synchronization id field of filters table. As we observed the values, the synchronization identity number was 1 for place, then for tweet id we observed that its value is 2, for time value we observed that its value is 3 and for date value was 4, and status was 5 all of these set values is stored and arranged in series order on the same column with field name ‘token’. We applied the DELETE and WILDCARD SQL command to delete the tweet identity number from token field due to its irrelevant to other values. As we performed the data migration task, we migrate the time values to its specific database table with the name ‘table’ using SELECT, WILDCARD and INSERT SQL Command. During the storing process, we used the formula $TmeSynId = Ts - 2$ where *TmeSynId* represents the synchronization id that is allocated in database time table and *Ts* represents the synchronization id value of time value when it was in token field of filters table before the migration was performed. The minus 2 represents the displacement value from synchronization id number of place. Secondly we transferred the status values to its specific table with the name of ‘status’ using the same SQL command that we used during first data migration. As storing process performed, we used the formula $StatSynId = Ss - 4$ where *StatSynId* represents the synchronization id value in status table, Ss represents the synchronization id value of staus value when it was in token field of filters table, the minus 4 represents the displacement value from synchronization id number of place. The last data migration we performed is the date values to its specific database date table using the same SQL command. During the storing process, we used the formula DteSynId = Ds – 3 where DteSynId represents the synchronization id value in date table, Ds represents the synchronization id value of date value when it was in token field of filters table before the migration. During the data migration was performed, The Three formulas generated the same synchronization id value of place from filters table that are distributed to time, status and date table the we applied LEFT JOIN , WHERE clause, and INSERT SQL command to three table and we acquired the data matrix that to be used for data mining.

4.3.4 – Data Preprocessing Steps

4.3.4.1 Data Preprocessing using List Wise Deletion Sequential Method

The data matrix we acquired experienced the missing value problem we had these due to merging of the three tables. To solve this situation we adopted sequential method of data preprocessing and used List wise deletion approach or also known as case wise deletion. In this approach, all cases with missing attribute values are deleted from the data set. After deletion, we acquired 12,375 instances with 4 attributes.

4.3.4.1 Filtering using Frequent Values Deletion

The data matrix we acquired experienced a frequent value in terms of traffic status.

Table 1 *Count of Traffic Status before the filtering method was applied*

status	Count
MH	4,629
M	8,101
L	5

Table 1 shows the count for Moderate to heavy Traffic status was 4,629, while Moderate status was 8,101 and Low status was 5

Table 2 *Count of Traffic Status before the filtering method was applied*

status	Count
MH	2,490
M	3,894
L	0

Table 2 shows that when filtering method was applied, the 12,735 instances reduced to 6, 384. Low traffic status was deleted and become 0 of the count, moderate to heavy traffic status reduced from 4,629 to 2, 490 and Moderate traffic status reduced from 8,101 to 3,894.

4.4 Data Modeling

We use J48 classification algorithm to determine the precision, recall and F-Measure of day, traffic status, time and place, which is given from data matrix. In this model, we can able to identify how exactly the time where routed roads will be experienced the moderate or moderate to heavy traffic

Table 3 *Precision, Recall and F-Measure of each day where traffic will be experienced*

Day	Precision	Recall	F Measure
Tuesday	0.999	1	0.999
Friday	1	0.997	0.999

Table 3 shows the precision, recall and f-measure of Tuesday and Friday where traffic will be experienced by commuters in Metro Manila, with the accuracy of 99.9%.

Table 4 *Precision, Recall and F-Measure of time traffic where Moderate and Moderate to Heavy traffic status will be experienced by NCR Cross Sectional Roads*

Time of Traffic	Precision	Recall	F Measure
8:09AM	1	0.6	0.75
8:46AM	1	0.6	0.75
8:47AM	1	0.8	0.88
9:41AM	0.75	0.8	0.77
1:15PM	0.75	1	0.86
1:17PM	1	1	1
1:38PM	1	0.68	0.8
1:41PM	1	1	1
1:42PM	1	1	1
2:30PM	1	0.8	0.89
2:31PM	1	1	1
2:34PM	1	1	1
2:40PM	1	0.6	0.75
2:44PM	1	1	1
2:48PM	1	0.938	0.968
2:52PM	1	1	1
2.55PM	1	1	1
3:07PM	1	0.83	0.90
3:09PM	1	0.85	0.92
3:11PM	1	1	1
3:12PM	0.97	0.76	0.85
3:14PM	1	0.75	0.85
3:44PM	0.92	0.62	0.74
4.07PM	1	1	1
4:13PM	1	0.80	0.88
4:22PM	1	1	1
4:38PM	1	1	1
4:41PM	1	1	1
6:12PM	0.75	1	0.86
6:14PM	1	1	1
6:15PM	0.91	0.84	0.88
8:00PM	1	0.6	0.75

Table 4 shows the precision, recall and f-measure of listed time where traffic will be experienced by commuters in Metro Manila, with the accuracy of 75.72 %.

Table 5 *Precision, Recall and F-Measure of Places that can experience moderate or moderate to heavy traffic at specific time listed from table 4*

Place	Precision	Recall	F Measure
C5 NB: Market Market - Libis	1	1	1
Ateneo to Miriam	0.50	1	0.66
Miriam to Xavaerville	0.50	1	0.66
Roxas Blvd: SB Anda - Finace	0.66	1	0.88
Bagong Ilog - Lanuza	1	0.68	0.8
Pedro Gil – UN Avenue	0.50	1	0.66
P. Noval - Lerma	1	1	1
Airport Road – R. Sulayman	0.6	1	0.75
C5 SB: T.Sora – C.P. Garcia	0.5	0.87	0.63
Kalayaan	0.50	0.87	0.63
A. Maceda – Lacson	0.50	1	0.66
Edsa NB: Harrison - Whiteplains	1	1	1
Niaroad – Shaw	1	1	1
Malibay - Megamall	1	1	1
B. Serrano - Kalayaan	0.50	0.91	0.646
Edsa NB: Harrison – Buendia	0.50	1	0.66
C5 NB: Market2 – B. Ilog	0.50	1	0.66
Lanuza - Libis	0.50	1	0.66
Timog – Main Avenue	0.97	0.76	0.85
Aurora - Ateneo	1	1	1
Miriam – CP Garcia	0.92	0.62	0.74

Table 5 shows the precision, recall and f-measure of routed road where traffic will be experienced by commuters in Metro Manila, with the accuracy of 45. 37%. Majority of routed roads is located in Edsa North Bound and South Bound and C5 North Bound and South Bound with Precision of 1.00, Recall of 1.00 and F- Measure of 1.00

5 CONCLUSIONS

The researcher concludes that during data transformation of semi structured MMDA tweets into structured data matrix several problem occur these are the tweet syntax inconsistency, trimming of dirty data while keeping other important data not to be destroyed, data replacing and merging to keep traffic time, traffic status, and day aligned to each other and the researcher also experienced the missing records. The researcher also proved that using several SQL command such as “SELECT”,”DELETE”, Wildcard for pattern matching “INSERT”, “WHERE”, “LIKE”, and “LEFT JOIN” can be used for extracting tweet with similar syntax and store it in a file and to be used later. The use of SQL command is also very useful in file merging to form a data matrix that can be used for data exploration and for sequential data preprocessing in handling missing value. The researcher also proved that using some of PHP string function is useful in MMDA tweet tokenization such as left trimming, string replacement, string reverse, explode function to transform the message into array. During the j48 algorithm is used we abled to enlist all traffic time where moderate and moderate to heavy can be experienced by the commuters from National capital Region. We discovered that from Eight O’clock to Nine Forty One in morning the commuters are experiencing a majority traffic status of moderate. We also discovered that from One Fifteen in the afternoon to Eight-O’clock in the evening the commuters are also experiencing the moderate and moderate to heavy traffic status in several routed roads in NCR. For the routed roads, we discovered that C5 North Bound and C5 South Bound are experiencing traffic every Tuesday and Friday in the afternoon and evening that comprises of routed roads. These are Market Market to Libis, Market Market to B. Ilog, and Tandang Sora to Carlos P. Garcia

ACKNOWLEDGMENT

The author wishes to thank Metropolitan Manila Development Authority (MMDA) in providing public information and to Dr. Menchita F. Dumlao for mentoring and encouraging in finishing this paper. This work was completed for the completion of the requirement in Advanced Data Mining subject.

REFERENCES

[1] C.S. George, "Economic Effects of Traffic in Metro Manila" http://www.businessmirror.com.ph/economic-effects-of-traffic-in-metro-manila/. 2015.

[2] "Stress, pollution, fatigue: How traffic jams affect your health" http://www.apastyle.org/learn/faqs/web-page-no-author.aspx. 2015.

[3] K. George, "What causes Fatigue? 251 Causes" http://www.healthline.com/symptom/fatigue.

[4] J.A Anne, "Cardiovascular Disease is still the country's top killer" http://lifestyle.inquirer.net/178609/cardiovascular-disease-is-still-the-countrys-top-killer/.2014

[5] M. Izabel, "What causes traffic jams in Metro Manila?" http://www.filipinoscribe.com/2015/09/05/what-causes-the-extreme-traffic-jams-in-metro-manila/. 2015

[6] "MMDA uses Twitter for Public Service" http://www.philstar.com/networks/711633/mmda-uses-twitter-public-service. 2011

[7] R. Wirth and J. Hipp, "CRISP: DM Towards a Standard Process Model for Data Mining " http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.198.5133&rep=rep1&type=pdf

[8] S. Robert, "Data is Useless without Meaning: The importance of insight" http://www.econtentmag.com/Articles/News/News-Feature/Data-Is-Useless-Without-Meaning-The-Importance-of-Insight-91693.htm. 2013.

[9] A. Iftikhar, K. Shah, R. Azhar, and Z. Qamruz, "Conducting Surveys and Data Collection: From Traditional Mobile and SMS –based Surveys" Pak.j.stat.oper.res. Vol. X No.2 2014 pp169 – 187, available at http://search.proquest.com/openview/920ad7e2cba14988f62f0ec260989f97/1?pq-origsite=gscholar.2014

[10] A.J. Stephen , "Business Impact of Web 2.0 Technologies" Communications of the ACM, Vol. 53 No. 12, Pages 67-79, available at http://cacm.acm.org/magazines/2010/12/102142-business-impact-of-web-2-0-technologies/fulltext.2010

[11] B. Lorenz, "The Most Important Employee Rights at the Workplace" https://www.salarium.com/important-employee-rights/. 2016

[12] G. Paul, "What Exactly Is 'Twitter'? What is 'Tweeting'?" https://www.lifewire.com/what-exactly-is-twitter-2483331. 2016.

[13] B. Kristian. "Understanding Sentiment Analysis: What it Is and Why It's Used" https://www.brandwatch.com/blog/understanding-sentiment-analysis/. 2015