Predictive Analytics

Test Before You Tweet

Team number

Anchal Gupta
Ishita Daga
Raj Vardhan

School of Graduate Professional Studies
Data Analytics
IE-575 – Predictive Analytics
Spring, 2019

1 DOCUMENT CONTROL

1.1 Work carried out by:

Name	Email Address	Task description
Anchal Gupta	aug1167@psu.edu	TF-IDF and classification models
Ishita Daga	ixd84@psu.edu	Doc2Vec and classification models
Raj Vardhan	rzv86@psu.edu	Data collection, exploration and Logistic Regression

1.2

1.3 Revision Sheet

Date	Revision Description
02/21/2019	Content revision
02/22/2019	Results addition
02/23/2019	References
02/24/2019	Formatting

TABLE OF CONTENTS

SECTION 1: INTRODUCTION

The connected society we live in today has allowed online users to willingly share opinions on an unprecedented scale. Motivated by the advent of mass opinion sharing, it is then crucial to devise algorithms that efficiently identify the emotions expressed within the opinionated content. Recently, Twitter has received a lot of interest and attention from a wide range of internet users across the globe. One of the main reasons for using Twitter is the ease of expressing opinions on diverse topics such as "Data Science". Such ease of use, coupled with the widespread use of connected portable devices, has made Twitter the primary channel for users to voluntarily share opinions, feelings, news, activities, interests, and other types of event-related information happening around them. Besides connecting with friends and sharing moments, social networks are used for a source of news and information. People share their views by posting their opinions in the form of Tweets. A basic fact is that different people are interested in different kinds of tweets, and they will retweet tweets and like other's tweets which they are interested in. Furthermore, showing user the content they prefer can increase their satisfaction.

SECTION 2: PROBLEM STATEMENT

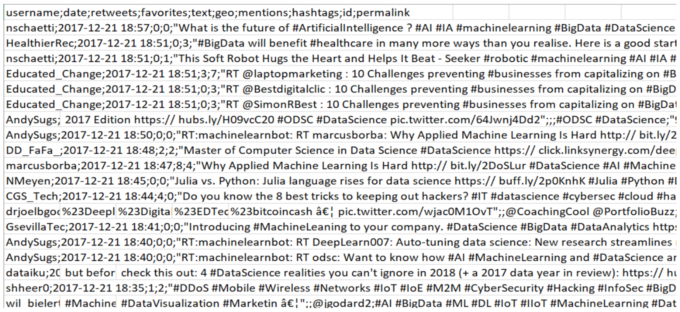
When a person writes a Tweet, it is expected that it's content will affect user and bring value to them. On Twitter, articles are posted including hashtags, URL's, titles etc. The character limit of the Tweet is 280 characters. Thus, to create good impression and enhance interest of a reader it is very important to focus on the content within the character limit. Users can access and express satisfaction in the form of likes and retweets(share) of the original post. Providing a prediction of number of likes and retweets on a post can help users to improvise their subject matter and engage more users.

SECTION 2.1: IMPORTANCE

Predictions of likes and retweets can improve the content of information we receive from people. It will also increase Twitter usage among readers to extract the valuable information. This in turn can be beneficial for enhancement in all types of business.

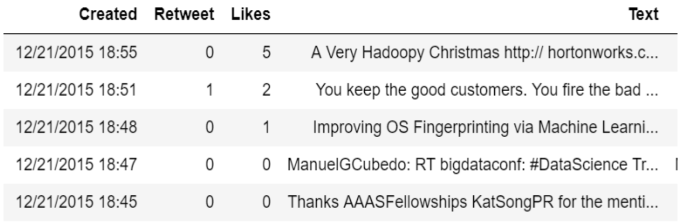
SECTION 3: DATA

We collected 2 million tweets from Twitter using tweepy library for the Year 2015 to 2018 in the field Data Science. We chose data science as it is the most trending domain among people from past few years. Following is the snapshot of the raw data collected.



```
username;date;retweets;favorites;text;geo;mentions;hashtags;id;permalink
nschaetti;2017-12-21 18:57;0;0;"What is the future of #ArtificialIntelligence ? #AI #IA #machinelearning #BigData #DataScience
HealthierRec;2017-12-21 18:51;0;3;"#BigData will benefit #healthcare in many more ways than you realise. Here is a good start
nschaetti;2017-12-21 18:51;0;1;"This Soft Robot Hugs the Heart and Helps It Beat - Seeker #robotic #machinelearning #AI #IA #
Educated_Change;2017-12-21 18:51;3;7;"RT @laptopmarketing : 10 Challenges preventing #businesses from capitalizing on #B
Educated_Change;2017-12-21 18:51;0;3;"RT @Bestdigitalclic : 10 Challenges preventing #businesses from capitalizing on #BigD
Educated_Change;2017-12-21 18:51;0;3;"RT @SimonRBest : 10 Challenges preventing #businesses from capitalizing on #BigDat
AndySugs;  2017 Edition https:// hubs.ly/H09vcC20 #ODSC #DataScience pic.twitter.com/64Jwnj4Dd2";;;#ODSC #DataScience;"
AndySugs;2017-12-21 18:50;0;0;"RT:machinelearnbot: RT marcusborba: Why Applied Machine Learning Is Hard http:// bit.ly/2
DD_FaFa_;2017-12-21 18:48;2;2;"Master of Computer Science in Data Science #DataScience https:// click.linksynergy.com/dee
marcusborba;2017-12-21 18:47;8;4;"Why Applied Machine Learning Is Hard http:// bit.ly/2DoSLur #DataScience #AI #Machine
NMeyen;2017-12-21 18:45;0;0;"Julia vs. Python: Julia language rises for data science https:// buff.ly/2p0KnhK #Julia #Python #I
CGS_Tech;2017-12-21 18:44;4;0;"Do you know the 8 best tricks to keeping out hackers? #IT #datascience #cybersec #cloud #ha
drjoelbgo %23Deepl %23Digita %23EDTec %23bitcoincash â€¦ pic.twitter.com/wjac0M1OvT";;@CoachingCool @PortfolioBuzz;
GsevillaTec;2017-12-21 18:41;0;0;"Introducing #MachineLeaning to your company. #DataScience #BigData #DataAnalytics http:
AndySugs;2017-12-21 18:40;0;0;"RT:machinelearnbot: RT DeepLearn007: Auto-tuning data science: New research streamlines
AndySugs;2017-12-21 18:40;0;0;"RT:machinelearnbot: RT odsc: Want to know how #AI #MachineLearning and #DataScience ar
dataiku;2C  but befor  check this out: 4 #DataScience realities you can't ignore in 2018 (+ a 2017 data year in review): https:// hu
shheer0;2017-12-21 18:35;1;2;"#DDoS #Mobile #Wireless #Networks #IoT #IoE #M2M #CyberSecurity #Hacking #InfoSec #BigD
wil_bielert #Machine #DataVisualization #Marketin â€¦";;@igodard2;#AI #BigData #ML #DL #IoT #IIoT #MachineLearning #Dat
```

We converted the data into data frame and removed unwanted columns like Username, geo, mentions, hashtags, id, permalink.



Created	Retweet	Likes	Text
12/21/2015 18:55	0	5	A Very Hadoopy Christmas http:// hortonworks.c...
12/21/2015 18:51	1	2	You keep the good customers. You fire the bad ...
12/21/2015 18:48	0	1	Improving OS Fingerprinting via Machine Learni...
12/21/2015 18:47	0	0	ManuelGCubedo: RT bigdataconf: #DataScience Tr...
12/21/2015 18:45	0	0	Thanks AAASFellowships KatSongPR for the menti...

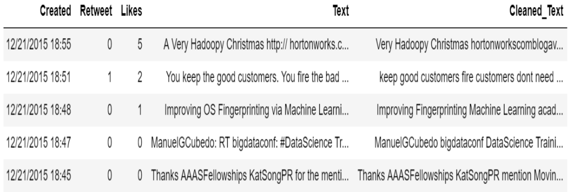
Feature	Description	Scale
Created	When the Tweet was posted	Timestamp
Retweet	How many times that Tweet was shared	Nominal
Likes	How many times that Tweet was liked	Nominal
Text	The content of the Tweet	String

SECTION 3.1: DATA PREPROCESSING

For the data preprocessing, we followed the below steps:

Step 1: Cleaning of Text Column

Tweet data had many unwanted characters and links. We used regular expressions in R to clean the column.



Created	Retweet	Likes	Text	Cleaned_Text
12/21/2015 18:55	0	5	A Very Hadoopy Christmas http:// hortonworks.c...	Very Hadoopy Christmas hortonworkscomblogav...
12/21/2015 18:51	1	2	You keep the good customers. You fire the bad ...	keep good customers fire customers dont need ...
12/21/2015 18:48	0	1	Improving OS Fingerprinting via Machine Learni...	Improving Fingerprinting Machine Learning acad...
12/21/2015 18:47	0	0	ManuelGCubedo: RT bigdataconf: #DataScience Tr...	ManuelGCubedo bigdataconf DataScience Traini...
12/21/2015 18:45	0	0	Thanks AAASFellowships KatSongPR for the menti...	Thanks AAASFellowships KatSongPR mention Movin...

Step 2: Stop words removal

We removed stop words (most frequent words with least significance for text procession like a, the, are etc.) using ntlk library.

Step 3: Removed Duplicates

We removed Duplicate Tweets after cleaning.

Step 4: Removed rows with 0 likes and retweets

We removed all the rows with having count of likes and retweets as 0.

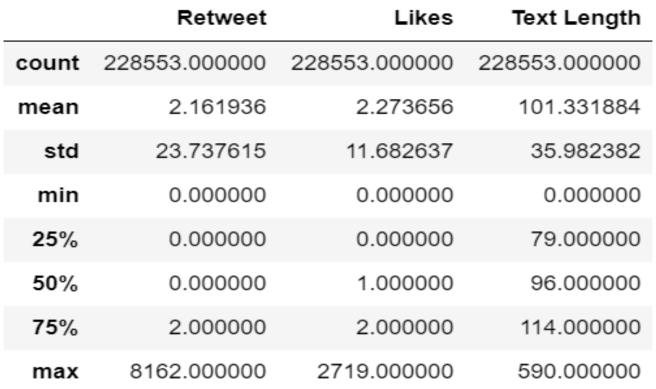
SECTION 4: METHODOLOGY

SECTION 4.1 EXPLORATORY VISUALIZATION

We explored the data to analyze the possible metrics that can be used to understand the solution. We identified the relationship between the features with overall performance of the tweet.

SECTION 4.1.1 OVERALL STATISTICS

We analyzed here the high-level statistics of the tweets and tried to understand how many times the tweets were on average retweeted or liked and the average length of a tweet.

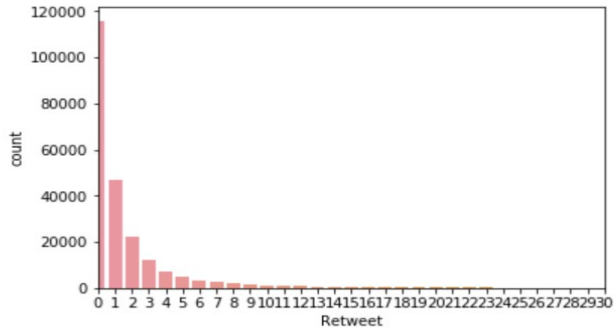


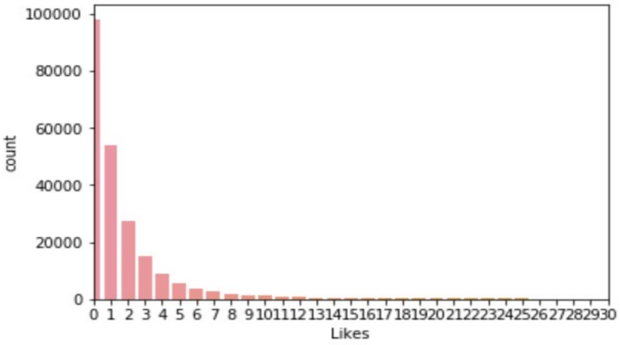
	Retweet	Likes	Text Length
count	228553.000000	228553.000000	228553.000000
mean	2.161936	2.273656	101.331884
std	23.737615	11.682637	35.982382
min	0.000000	0.000000	0.000000
25%	0.000000	0.000000	79.000000
50%	0.000000	1.000000	96.000000
75%	2.000000	2.000000	114.000000
max	8162.000000	2719.000000	590.000000

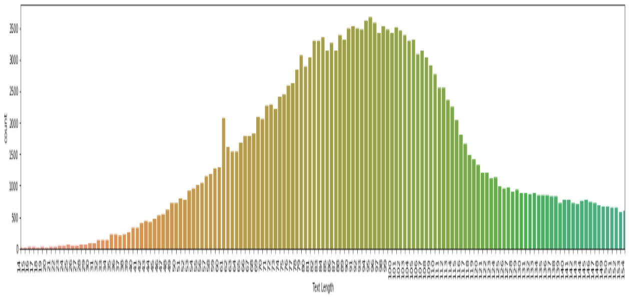
The above results helped us to understand the order of magnitude of our dataset. We could find out the minimum and maximum values of our attributes.

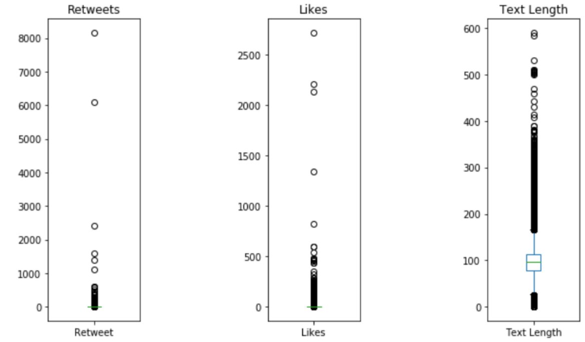
SECTION 4.1.2 HISTOGRAMS AND BOXPLOTS

In this section we will check how the multiple features are distributed. From these histograms, together with the overall statistic and the box plots, we can notice that the average length is around 90 characters. Likes and retweets are positive-skewed, i.e. they are concentrated on the left part of the graph.









To avoid being biased by outliers, we removed for each feature analysis the data points that don’t fit the following formulas:

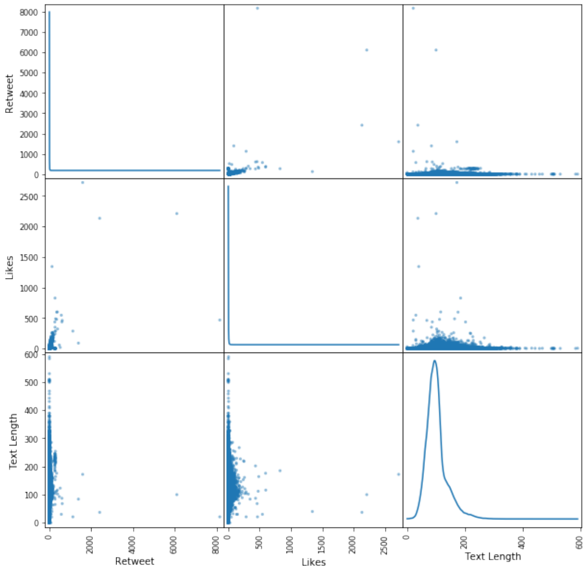
$$Outlier < Q1 - 3*IQR$$

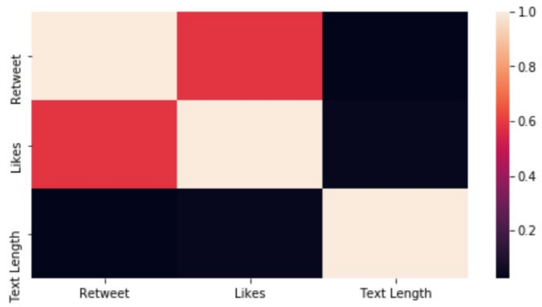
$$Outlier > Q3 + 3.5*IQR$$

Where Q1 and Q3 are the first and third quartile and IQR is the Interquartile Range (IQR = Q3 - Q1)

SECTION 4.1.3 SCATTER MATRIX AND CORRELATIONS

We tried to find a relationship between the multiple features. We could observe that there is a correlation between Likes and Retweets.

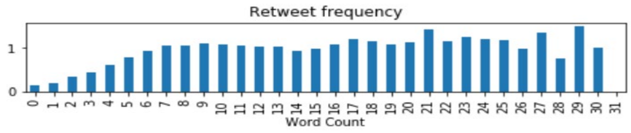


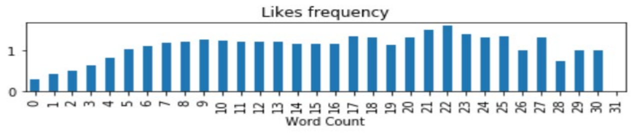


We removed Text length as it has no correlation with Likes and retweets.

SECTION 4.1.4 PERFORMANCE OF NUMBER OF WORDS

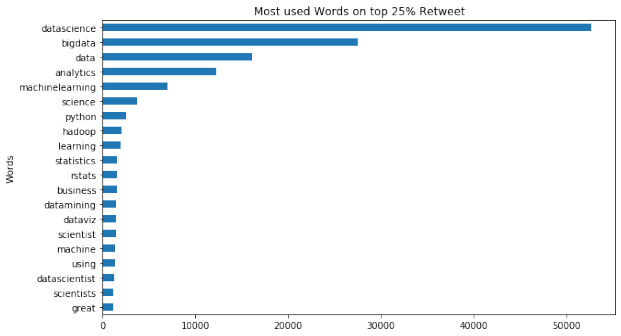
We analyzed the word count and reached the conclusion that the best number of words in the tweet is from 9 to 29 words.

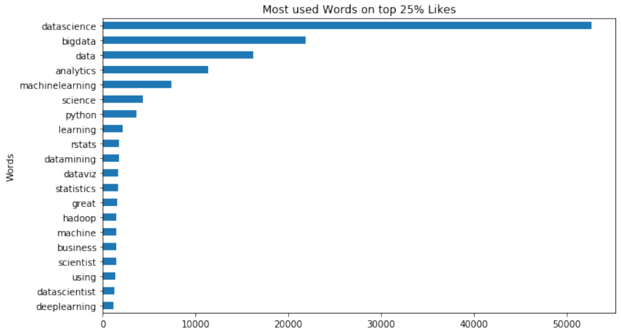




SECTION 4.1.5 WORDS THAT PERFORMED BETTER

We analyzed the top 25% performers for each one of the features to observe wat type of words are mostly used by people.

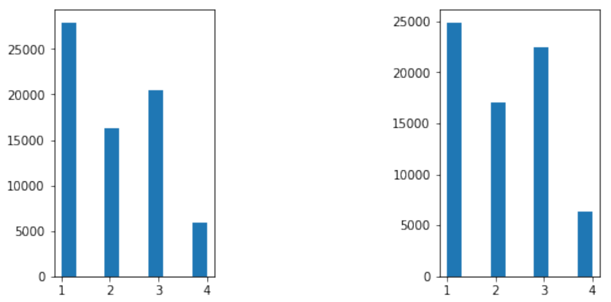




SECTION 4.2 BINNING OF TARGET VARIABLE

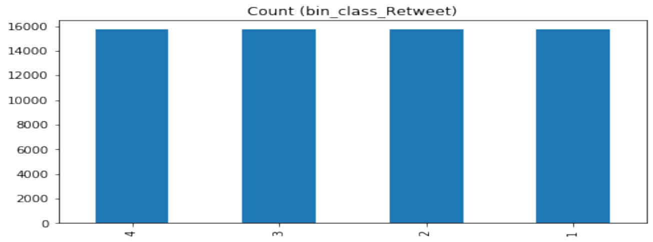
SECTION 4.2.1 DIVISION OF DATA INTO BINS

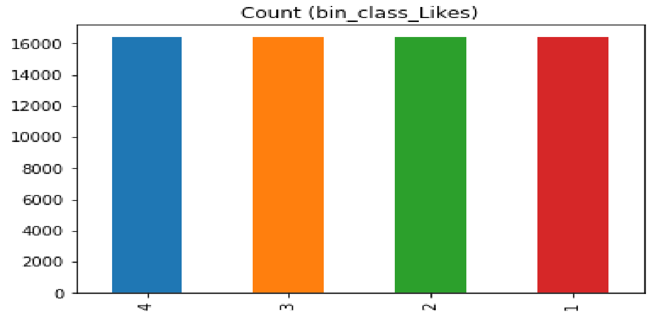
We binned our data into 4 classes:
1: [0-1], 2: [1-2], 3: [2-5], 4: [5+]



SECTION 4.2.2 DOWN SAMPLING OF DATA

Due to imbalanced data, we performed down sampling to have equal number of data points in each class.





SECION 4.3 ALGORITHMS AND TECHNIQUES

SECTION 4.3.1 BAG OF WORDS

To analyze the tweets in each data point, we need to map each word into a number. This is necessary because machine learning models normally don't process raw text, but numerical values. To reach this, we used a bag of words model. In this model, it is taken into consideration the presence and often the frequency of words, but the order or position is ignored. For the calculation of the bag of words, we will use a measure called Term Frequency, Inverse Document Frequency (TF-IDF) and Doc2Vec (Document to Vector). The goal is to limit the impact of tokens (words) that occur very frequently. We created a Pipeline which applied Count Vectorizer, TF-IDF and Models in sequence.

SECTION 4.3.2 TRAINING AND TESTING DATA SPLIT

Before starting the training and the evaluation of the models, we split the dataset into test and training sets. Retweets and lik es to have a total of 62856 data points each after all cleaning, with 43999(70%) as training and 18857 (30%) as testing point s.

SECTION 4.3.3 MODELS

Classification is a common task of machine learning which involves predicting a target variable taking into consideration the previous data. To reach such classification, it is necessary to create a model with the previous training data, and then use it to predict the value of the test data. This process is called supervised learning, since the data processing phase is guided toward the class variable while building the model. Predicting the number of retweets and likes of an article can be treated as a classification problem, because the output will be discrete values (range of numbers). We applied the following classification models to our dataset:

SECTION 4.3.3.1 LOGISTICS REGRESSION

The Logistic regression estimates the parameters of this function (coefficients), and as result it predicts the probability of presence of the characteristic of interest. This model was chosen, because provides probabilities for outcomes and a convenient probability scores for observations.

SECTION 4.3.3.2 SVM WITH LINEAR KERNEL

SVM constructs a hyperplane (or a set) that can separate the points in the defined labels. The distance between the closest data points and the hyperplane is named margin. An ideal separation is defined by a hyperplane that has the largest distance to the closest points of any class, so the challenge is to find the coefficients that maximize this margin. This model was chosen, because it works well with linear and non-linear datasets. And due the fact we have more samples than number of features, it can generate a good prediction.

SECTION 4.3.3.3 RANDOM FOREST

Random forest is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. We used this model to avoid overfitting of data.

SECTION 4.3.3.4 NEURAL NETWORK

Deep Neural Networks are more complex neural networks in which the hidden layers perform much more complex operations than simple sigmoid or relu activations. We used this network as we had large number of data points.

SECTION 4.3.3.5 MULTINOMIAL NAÏVE BAYES

The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as TF-IDF may also work.

SECTION 5: RESULTS

For each model tested, we calculated the accuracy for the default model (without any parameter, just the default ones) and, we tried to come with better parameters to evolve the accuracy. To test the combination of the new parameters, fine tune the model, we used grid search (GridSearchCV). To estimate the model's accuracy, we used a 5-fold cross validation that split the dataset into 5 parts, 4 of training and 1 of testing.

SECTION 5.1 ACCURACY RETWEETS (%)

Model Name	**TF-IDF**		**Doc2Vec**	
	Train	Test	Train	Test
Logistic Regression	47.6	49.03	37.03	35.68
Support Vector Machine	51.5	55.96	32.2	35.4
Random Forest	51.04	54.5	80.02	62.67
Neural Network	44.6	39.10	54.16	44.3
Multinomial Naïve Bayes	51.4	56.22	34.02	35.68

SECTION 5.2 ACCURACY LIKES (%)

Model Name	**TF-IDF**		**Doc2Vec**	
	Train	Test	Train	Test
Logistic Regression	47.6	49.03	36.20	35.06
Support Vector Machine	52.6	55.2	35.9	34.8
Random Forest	51.4	54.8	90.28	60.08
Neural Network	49.18	42.77	47.35	40.65
Multinomial Naïve Bayes	48.4	50.5	33.38	35.08

SECTION 5.3 DISCUSSION OF RESULTS

Multinomial Naïve Bayes performed best with accuracy of 56.22% in case of Retweets and Support Vector Machine performed best with accuracy of 55.2% in case of Likes. Random forest surprisingly overfitted the data in both the cases.

SECTION 6: REFERENCES

1.https://www.twitter.com.
2.https://www.medium.com.
3.https://medium.freecodecamp.org/.
4.N. Abdelhamid, A. Ayesh, F. Thabtah, S. Ahmadi, W. Hadi. MAC: A multiclass associative classification algorithm J. Info. Know. Mgmt. (JIKM), 11 (2) (2012), pp. 125001-1-1250011-10 WorldScinet.
5.I.H. Witten, E. Frank, M.A. Hall. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington, MA (2011).
6. Thabtah, S. Hammoud, H. Abdeljaber, Parallel associative classification data mining frameworks based mapreduce, To Appear in Journal of Parallel Processing Letter, March 2015, World Scientific, 2015.
7. http://www.numpy.org/.
8. http://scikit-learn.org/stable/.