Study Project based on the analysis of

VEWS: A Wikipedia Vandal Early Warning System

Graduate-Advisor: Dr. Aidong Lu

Aravind Reddy Keesara, 800976233
Ishan Agarwal, 800986801
Chandrakanth Yela, 800987193
Department of Computer Science University of North Carolina at Charlotte
Charlotte, North Carolina

Abstract

As the world's biggest information source, Wikipedia is being compromised by small number of vandals. This paper presents a system VEWS (Vandal Early Warning System) which is an effective approach for early identification of vandals on Wikipedia. Users are ideally classified as Benign and Vandal. Benign users are genuine users whereas Vandal users pose a threat to the information source. Vandal users can be identified and presented early to Wikipedia administrators. VEWS identifies such incidents before any human or known vandalism detection system. In order to perform the analysis, we use UMD Wikipedia Dataset.

As a part of group project, Extracting the Wikipedia Dataset and Sampling the dataset to perform the analysis are the common part for each member. Features of Vandal Users are identified using the Wikipedia Vandal Behavior approach in which the editing pattern of users are used as a criteria. This approach captures edit summary of the users. For effective visualization of these features, we can use Business Intelligence tool like Tableau. Final approach is VEWS, a combination of both the approaches, which has higher accuracy value in the prediction of vandal users.

I. INTRODUCTION

Wikipedia has been compromised by individuals who performs acts of vandalism and this vandalism has been defined by Wikipedia as “Manipulation of contents which compromises integrity of Wikipedia”. Wikipedia has taken necessary actions in eradicating the vandalism that has been taking place. In an attempt, many detection techniques were being developed like ClueBot NG, STiki and Snuggle. Each detection system has its own detection technique implementation. ClueBot NG uses artificial neural network which scores edits and depending upon the score that the edit receive, the edit will be reverted if received worst score by the ClueBot. STiki is another tool which helps genuine users based upon the users reputation score and helps them to revert the vandalized contents, it uses edit Meta data. Snuggle uses heuristic rules and machine learning algorithm to flag acts of vandalism. Though there are many approaches in detecting vandalism, an efficient approach/algorithm has to be developed which detects the vandalism even before it happens and reports it to the Wikipedia administrators. Thus came the idea of VEWS which detects faster and more vandals when compared with ClueBot NG. Combination of ClueBot NG approach and VEWS would produce higher accuracy in detecting the vandals.

II. DATASET

A. *About Dataset*

The dataset that is being used for the vandal detection is the data collected at a time period of 19 months from Jan 2013 to July 2014 and the data is divided into two sets one to be Benign users and other Vandal users. Out of all users 16549 users are being identified as benign users and 17027 users are blocked as vandal users. Data files are available in formats of csv and tsv.

B. *Files Description*

Users.csv: users data consists of all registered users in Wikipedia and below are the attributes for a user:

- username: It is the name of the registered user
- userid: It is the unique identifier for the registered user
- Blocked_time**:** This timestamp corresponds to time when user was blocked by the administrators
- Blocked_reason: This specifies the reason for which the administrator has blocked the user
- Type: It indicates the type of user, whether benign or vandal

Pages.csv: This file consists of information related to the pages that are edited by users which are available in user dataset.

- Pagetitle: It is the title for the page
- Pageid: It is the unique identifier of the page
- Pagecategories: It is the categories that is assigned for the page by Wikipedia

benign_<year>_<month>.csv**:** benign users data are stored year month wise and edits are sorted chronologically and below are the attributes for benign users

- username: It is the unique identifier of the user

- revid: It is the id of the revision made by the user
- revtime : It is the time the revision is made by the user
- pagetitle: It provides information regarding title of the page that was edited
- isReverted: It indicates whether the revision is reverted by other user/bot in future
- revertTime: It specifies the time when the revision is reverted
- cluebotRevert: It indicates whether ClueBot NG reverts the edit
- stiki_score: It is the score for revision given by STiki
- stiki_REP_USER: It is the user reputation score given after the revision by STiki

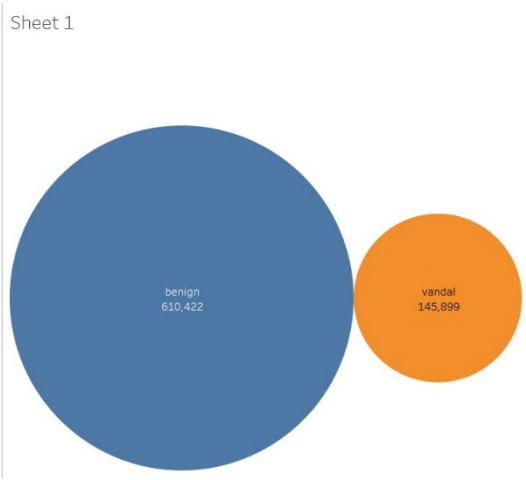
vandal_<year>_<month>.csv**:** This files consists of edits made by users who are blocked by Wikipedia administrators. Attributes are the same that are described for benign users.

- username: It is the unique identifier of the user
- revid: It is the id of the revision made by the user
- revtime : It is the time the revision is made by the user
- pagetitle: It provides information regarding title of the page that was edited
- isReverted: It indicates whether the revision is reverted by other user/bot in future
- revertTime: It specifies the time when the revision is reverted
- cluebotRevert: It indicates whether ClueBot NG reverts the edit
- stiki_score: It is the score for revision given by STiki
- stiki_REP_USER: It is the user reputation score given after the revision by STiki

IV. Preliminary visualization of data

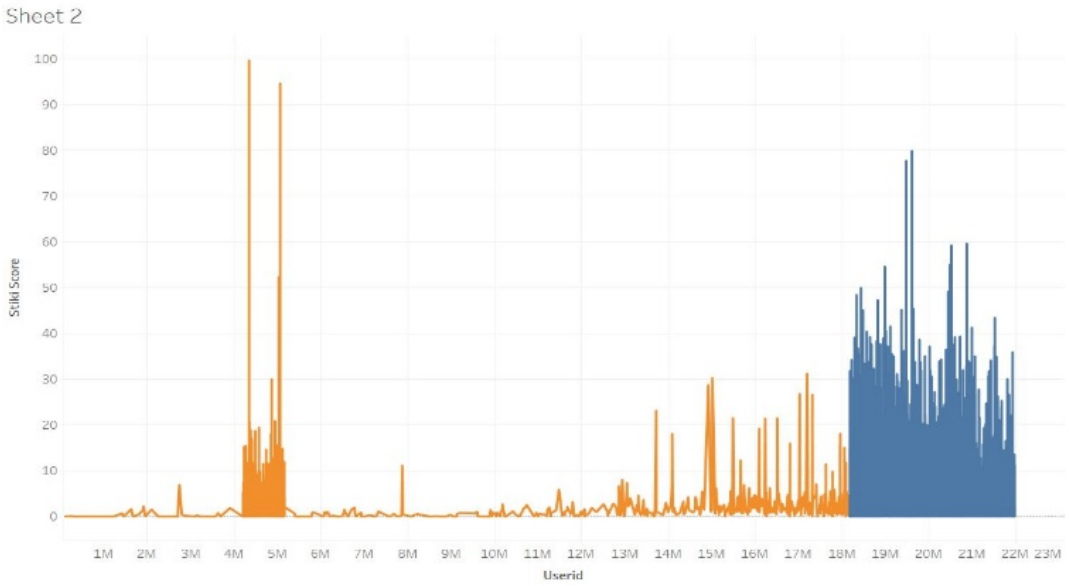
This part of analysis is done to get an overall idea on the UMDWikipedia dataset. The ".csv "file that we have obtained after cleaning data can be loaded into Tableau for exploring the data visually. Tableau is Business Intelligent Tool, which helps to gain the insights and understand data just by visualizing the statistics of the data. Tableau is capable of performing statistical analysis on large datasets. The generated csv file is loaded into the tableau.

A. Benign and Vandal users: For the given data set from 2013 to 2014, we may find the detection of vandal users by vandal detecting systems.



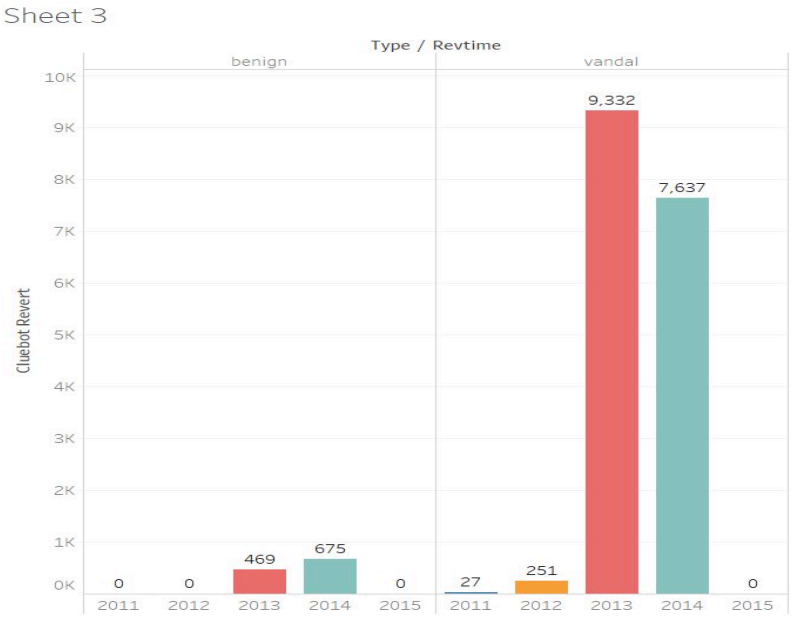
Benign vs Vandal

B. STiki score for Vandal and Benign users: STiki score for user reputations for benign and vandal can be seen clearly. Blue indicates benign users and orange indicates vandal users.



STiki Score

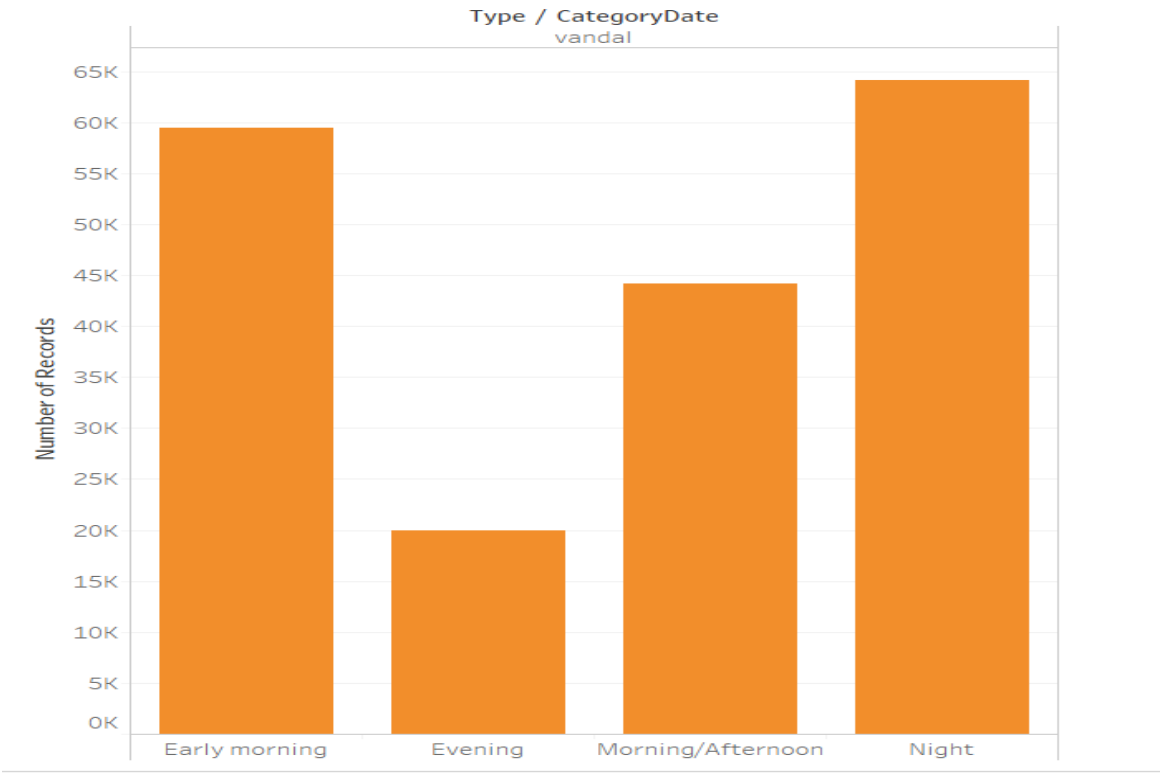
C. ClueBot NG for Vandal and Benign Users: ClueBot reverts from 2011 to 2015 can be seen for benign on left and vandal users on right of the sheet



ClueBot NG

Above are some of the visual analysis of the imported data in the tableau. Several other features are need to be explored by using ClueBot NG and VEWS analysis in detecting the vandalism before any other non-human detecting system.

D. The order in which the vandalism behavior varies as they are “Hyper Active” in Night, “Highly Active” in Early morning, “Moderately Active” in Morning/Afternoon, “Slightly Active” in Evening. This can be depicted by following visualization.



III. IMPLEMENTATION

For performing data analysis languages like R, Python can be used. The data set that is considered for vandal detection is large set, Python would be best choice as large datasets can be dealt easily by it. Ipython or Jupyter notebook is an interactive IDE to perform python operations. We can perform command line operations using jupyter.

Study Project - Team members : Chandrakanth Yela, Ishan Agarwal, Aravind Reddy Keesara

Importing packages

In [950]:

```
import pandas as pd
import numpy as np
import time
```

Reading Benign Data

In [951]:

```
db1 = pd.read_csv('benign_2013_1.csv')
db1['year']='January 2013'
db2 = pd.read_csv('benign_2013_2.csv')
db2['year']='February 2013'
db3 = pd.read_csv('benign_2013_3.csv')
db3['year']='March 2013'
db4 = pd.read_csv('benign_2013_4.csv')
db4['year']='April 2013'
db5 = pd.read_csv('benign_2013_5.csv')
db5['year']='May 2013'
db6 = pd.read_csv('benign_2013_6.csv')
db6['year']='June 2013'
db7 = pd.read_csv('benign_2013_7.csv')
db7['year']='July 2013'
db8 = pd.read_csv('benign_2013_8.csv')
db8['year']='August 2013'
db9 = pd.read_csv('benign_2013_9.csv')
db9['year']='September 2013'
db10 = pd.read_csv('benign_2013_10.csv')
db10['year']='October 2013'
db11 = pd.read_csv('benign_2013_11.csv')
db11['year']='November 2013'
db12 = pd.read_csv('benign_2013_12.csv')
db12['year']='December 2013'
db13 = pd.read_csv('benign_2014_1.csv')
db13['year']='January 2014'
db14 = pd.read_csv('benign_2014_2.csv')
db14['year']='February 2014'
db15 = pd.read_csv('benign_2014_3.csv')
db15['year']='March 2014'
db16 = pd.read_csv('benign_2014_4.csv')
db16['year']='April 2014'
db17 = pd.read_csv('benign_2014_5.csv')
db17['year']='May 2014'
db18 = pd.read_csv('benign_2014_6.csv')
db18['year']='June 2014'
db19 = pd.read_csv('benign_2014_7.csv')
db19['year']='July 2014'
```

```
db19['year'] = 'July 2014'
```

Merging the benign dataset

In [952]:

```
combined = db1
combined = combined.append([db2,db3,db4,db5,db6,db7,db8,db9,db10,db11,db12,
db13,db14,db15,db16,db17,db18,db19])
combined=pd.DataFrame(combined)
combined.head()
```

Out[952]:

	username	revid	revtime	pagetitle	isReverted	revertTime	clueb
0	Swiftscw	535365129	2013-01-28T16:43:56Z	User:Swiftscw	False	-	0
1	Swiftscw	535365309	2013-01-28T16:45:08Z	User:Swiftscw	False	-	0
2	Swiftscw	535366071	2013-01-28T16:50:17Z	User:Swiftscw	False	-	0
3	Swiftscw	535369287	2013-01-28T17:13:40Z	User:Swiftscw/AMAPOLA CABASE	False	-	0
4	Swiftscw	535370277	2013-01-28T17:21:30Z	Wikipedia:Requests for undeletion	False	-	0

Reading user file

In [953]:

```
colnames = ['username','userid','blocked_time','blocked_reason','type']
type(colnames)
users=pd.read_table('users.tsv',sep="\t",header=None,names=colnames)
users=pd.DataFrame(users)
users.head()
```

Out[953]:

	username	userid	blocked_time	blocked_reason	type
0	Swiftscw	18342896	-	-	benign
1	Azkle	18359941	-	-	benign
2	Tgrosinger	18228795	-	-	benign
3	BanglarBogra	18197517	-	-	benign
4	Tiny Trot	18315964	-	-	benign

Benign and User Left Join on "username"

In [954]:

```
ben = pd.merge(combined,users,how='left',on='username')
ben.head(10)
```

Out[954]:

	username	revid	revtime	pagetitle	isReverted	revertTime	clueb
0	Swiftscw	535365129	2013-01-28T16:43:56Z	User:Swiftscw	False	-	0
1	Swiftscw	535365309	2013-01-28T16:45:08Z	User:Swiftscw	False	-	0
2	Swiftscw	535366071	2013-01-28T16:50:17Z	User:Swiftscw	False	-	0
3	Swiftscw	535369287	2013-01-28T17:13:40Z	User:Swiftscw/AMAPOLA CABASE	False	-	0
4	Swiftscw	535370277	2013-01-28T17:21:30Z	Wikipedia:Requests for undeletion	False	-	0
5	Swiftscw	535376382	2013-01-28T18:07:16Z	Maria Cabase	False	-	0
6	Swiftscw	535379254	2013-01-28T18:27:12Z	Maria Cabase	False	-	0
7	Swiftscw	535380411	2013-01-28T18:34:51Z	Maria Cabase	False	-	0
8	Swiftscw	535382439	2013-01-28T18:48:17Z	Maria Cabase	False	-	0
9	Swiftscw	535390355	2013-01-28T19:40:51Z	Maria Cabase	False	-	0

Reading Vandal data

In [955]:

```
dv1 = pd.read_csv('vandal_2013_1.csv')
dv1['year']='January 2013'
dv2 = pd.read_csv('vandal_2013_2.csv')
dv2['year']='February 2013'
dv3 = pd.read_csv('vandal_2013_3.csv')
dv3['year']='March 2013'
dv4 = pd.read_csv('vandal_2013_4.csv')
dv4['year']='April 2013'
dv5 = pd.read_csv('vandal_2013_5.csv')
dv5['year']='May 2013'
dv6 = pd.read_csv('vandal_2013_6.csv')
dv6['year']='June 2013'
```

```
dv7 = pd.read_csv('vandal_2013_7.csv')
dv7['year']='July 2013'
dv8 = pd.read_csv('vandal_2013_8.csv')
dv8['year']='August 2013'
dv9 = pd.read_csv('vandal_2013_9.csv')
dv9['year']='September 2013'
dv10 = pd.read_csv('vandal_2013_10.csv')
dv10['year']='October 2013'
dv11 = pd.read_csv('vandal_2013_11.csv')
dv11['year']='November 2013'
dv12 = pd.read_csv('vandal_2013_12.csv')
dv12['year']='December 2013'
dv13 = pd.read_csv('vandal_2014_1.csv')
dv13['year']='January 2014'
dv14 = pd.read_csv('vandal_2014_2.csv')
dv14['year']='February 2014'
dv15 = pd.read_csv('vandal_2014_3.csv')
dv15['year']='March 2014'
dv16 = pd.read_csv('vandal_2014_4.csv')
dv16['year']='April 2014'
dv17 = pd.read_csv('vandal_2014_5.csv')
dv17['year']='May 2014'
dv18 = pd.read_csv('vandal_2014_6.csv')
dv18['year']='June 2014'
dv19 = pd.read_csv('vandal_2014_7.csv')
dv19['year']='July 2014'
```

Merging the Vandal dataset

In [956]:

```
comvandal = dv1
comvandal = comvandal.append([dv2,dv3,dv4,dv5,dv6,dv7,dv8,dv9,dv10,dv11,dv12,dv13,dv14,dv15,dv16,dv17,dv18,dv19])
comvandal=pd.DataFrame(comvandal)
comvandal.head()
```

Out[956]:

	username	revid	revtime	pagetitle	isReverted	revertTime	cluebotRevert	st
0	Alex12rocks	535937612	2013-01-31T23:19:17Z	Pueblo Revolt	True	2013-01-31T23:20:35Z	0	0.
1	Alex12rocks	535937732	2013-01-31T23:20:04Z	Pueblo Revolt	True	2013-01-31T23:20:35Z	0	0.
2	Alex12rocks	535937933	2013-01-31T23:21:19Z	Pueblo Revolt	True	2013-01-31T23:21:30Z	0	0.
3	Alex12rocks	535938394	2013-01-31T23:24:08Z	Pueblo Revolt	True	2013-01-31T23:24:15Z	0	0.
4	Alex12rocks	535938540	2013-01-31T23:24:56Z	Pueblo Revolt	True	2013-01-31T23:24:59Z	1	0.

Vandal and User Left Join on "username"

In [957]:

```
van = pd.merge(comvandal,users,how='left',on='username')
van.head()
```

Out[957]:

	username	revid	revtime	pagetitle	isReverted	revertTime	cluebotRevert	st
0	Alex12rocks	535937612	2013-01-31T23:19:17Z	Pueblo Revolt	True	2013-01-31T23:20:35Z	0	0.
1	Alex12rocks	535937732	2013-01-31T23:20:04Z	Pueblo Revolt	True	2013-01-31T23:20:35Z	0	0.
2	Alex12rocks	535937933	2013-01-31T23:21:19Z	Pueblo Revolt	True	2013-01-31T23:21:30Z	0	0.
3	Alex12rocks	535938394	2013-01-31T23:24:08Z	Pueblo Revolt	True	2013-01-31T23:24:15Z	0	0.
4	Alex12rocks	535938540	2013-01-31T23:24:56Z	Pueblo Revolt	True	2013-01-31T23:24:59Z	1	0.

Reading Page Details Data

In [958]:

```
pages=pd.read_csv('pages.tsv')
pages=pd.DataFrame(pages)
pages.head(10)
```

Out[958]:

	pagetitle	pageid	pagecategories
0	User:Swiftscw	38354683	set([u'Category:WikiProject Biography particip...
1	User:Swiftscw/AMAPOLA CABASE	38354972	set([])
2	Wikipedia:Requests for undeletion	26803395	set([u'Category:Wikipedia maintenance', u'Cate...

	pagetitle	pageid	pagecategories
3	Maria Cabase	32754928	set([u'Category:Filipino pianists', u'Category...
4	Talk:Maria Cabase	34549656	set([u'Category:WikiProject Biography articles...
5	Manuel Kabajar Cabase	32755116	set([u'Category:Filipino drummers', u'Category...
6	User talk:Postdlf/Archive23	38090775	set([])
7	Wikipedia:WikiProject Biography/Members	6163553	set([])
8	User talk:Swiftscw	40011085	set([])
9	Bobby Enriquez	7040405	set([u'Category:Filipino Pentecostals', u'Cate...

Left Join Vandal and Benign with Page data on "pagetitle"

In [959]:

```
van_final = pd.merge(van,pages,how='left',on="pagetitle")
ben_final = pd.merge(ben,pages,how='left',on="pagetitle")
van_final.head(10)
ben_final.head(10)
```

Out[959]:

	username	revid	revtime	pagetitle	isReverted	revertTime	clueb
0	Swiftscw	535365129	2013-01-28T16:43:56Z	User:Swiftscw	False	-	0
1	Swiftscw	535365309	2013-01-28T16:45:08Z	User:Swiftscw	False	-	0
2	Swiftscw	535366071	2013-01-28T16:50:17Z	User:Swiftscw	False	-	0
3	Swiftscw	535369287	2013-01-28T17:13:40Z	User:Swiftscw/AMAPOLA CABASE	False	-	0
4	Swiftscw	535370277	2013-01-28T17:21:30Z	Wikipedia:Requests for undeletion	False	-	0
5	Swiftscw	535370277	2013-01-28T17:21:30Z	Wikipedia:Requests for undeletion	False	-	0
6	Swiftscw	535376382	2013-01-28T18:07:16Z	Maria Cabase	False	-	0
7	Swiftscw	535379254	2013-01-28T18:27:12Z	Maria Cabase	False	-	0
8	Swiftscw	535380411	2013-01-	Maria Cabase	False	-	0

	username	revid	28T18:34:51Z revtime	pagetitle	isReverted	revertTime	clueb
9	Swiftscw	535382439	2013-01-28T18:48:17Z	Maria Cabase	False	-	0

Cleaning the data

In [960]:

```
import pandas as pd
```

In [961]:

```
van_final.apply(lambda x: sum(x.isna()))
```

Out[961]:

```
username            0
revid               0
revtime             0
pagetitle           0
isReverted          0
revertTime          0
cluebotRevert       0
stiki_score         0
stiki_REP_USER    727
year                0
userid              0
blocked_time        0
blocked_reason      0
type                0
pageid              0
pagecategories      0
dtype: int64
```

In [962]:

```
users_van = van_final.dropna(axis=0)
users_van.apply(lambda x: sum(x.isna()))
```

Out[962]:

```
username          0
revid             0
revtime           0
pagetitle         0
isReverted        0
revertTime        0
cluebotRevert     0
stiki_score       0
stiki_REP_USER    0
year              0
userid            0
blocked_time      0
blocked_reason    0
type              0
pageid            0
pagecategories    0
dtype: int64
```

In [963]:

```
users_ben = ben_final.dropna(axis=0)
users_ben.apply(lambda x: sum(x.isna()))
```

Out[963]:

```
username          0
revid             0
revtime           0
pagetitle         0
isReverted        0
revertTime        0
cluebotRevert     0
stiki_score       0
stiki_REP_USER    0
year              0
userid            0
blocked_time      0
blocked_reason    0
type              0
pageid            0
pagecategories    0
dtype: int64
```

Merging Vandal and Benign Data

In [964]:

```
users_fin=users_van.append(users_ben)
users_fin.head()
```

Out[964]:

	username	revid	revtime	pagetitle	isReverted	revertTime	cluebotRevert	st
0	Alex12rocks	535937612	2013-01-31T23:19:17Z	Pueblo Revolt	True	2013-01-31T23:20:35Z	0	0.
1	Alex12rocks	535937732	2013-01-31T23:20:04Z	Pueblo Revolt	True	2013-01-31T23:20:35Z	0	0.
2	Alex12rocks	535937933	2013-01-31T23:21:19Z	Pueblo Revolt	True	2013-01-31T23:21:30Z	0	0.
3	Alex12rocks	535938394	2013-01-31T23:24:08Z	Pueblo Revolt	True	2013-01-31T23:24:15Z	0	0.

	username	revid	revtime	pagetitle	isReverted	revertTime	cluebotRevert	st
4	Alex12rocks	535938540	2013-01-31T23:24:56Z	Revolt	True	2013-01-31T23:24:59Z	1	0.

converting date to proper format

In [965]:

```
ben_final.revtime = ben_final.revtime.str[:19]
```

In [966]:

```
ben_final.shape
```

Out[966]:

```
(657863, 16)
```

In [967]:

```
van.revtime = ben_final.revtime.str[:19]
```

In [968]:

```
ben_final['revtime'] = pd.to_datetime(ben_final['revtime'])
```

In [969]:

```
van_final['revtime'] = pd.to_datetime(van_final['revtime'])
```

seperating some cloumns to process other features in the later part of analysis

In [970]:

```
ben_scores = ben_final.loc[:,['userid','stiki_score','cluebotRevert']]
van_scores = van_final.loc[:,['userid','stiki_score','cluebotRevert']]
```

1) checking whether the first edit of user is a metapage

In [971]:

```
ben_final[ben_final['pagetitle'].str.contains('/')]
```

Out[971]:

	username	revid	revtime	pagetitle
3	Swiftscw	535369287	2013-01-28 17:13:40	User:Swiftscw/AMAPOLA CABASE
16	Swiftscw	535402422	2013-01-28 21:00:26	User talk:Postdlf/Archive23

	username	revid	revtime	pagetitle
18	Swiftscw	564933523	2013-07-19 15:02:39	Wikipedia:WikiProject Biography/Members
26	Azkle	535906972	2013-01-31 19:50:17	Wikipedia:Training/For students/Training feedback
97	Tgrosinger	537167961	2013-02-08 05:00:23	User:Tgrosinger/sandbox
100	Tgrosinger	539104591	2013-02-19 20:42:51	Wikipedia talk:Articles for creation/Storm
106	Tgrosinger	539914351	2013-02-23 16:49:06	Wikipedia:WikiProject Seattle/Active participants
118	Tgrosinger	540982409	2013-02-27 16:42:58	Talk:Torchwood/GA1
119	Tgrosinger	540989343	2013-02-27 17:12:00	Talk:Torchwood/GA1
152	Naveenchand3	531288293	2013-01-04 15:55:16	Wikipedia:WikiProject Pakistan/Assessment
158	Griff88	533205575	2013-01-15 14:29:37	User:Griff88/Do svidanja, Moskva
159	Griff88	533206041	2013-01-15 14:34:15	User:Griff88/Do svidanja, Moskva
160	Griff88	533343359	2013-01-16 09:46:16	User:Griff88/Do svidanja, Moskva
161	Griff88	533343583	2013-01-16 09:49:54	User:Griff88/Do svidanja, Moskva
176	Griff88	550442023	2013-04-15 08:22:00	Wikipedia:WikiProject Multi-sport events/Asian...
186	Griff88	556398136	2013-05-23 08:49:05	Wikipedia:WikiProject Olympics/Members
268	Griff88	587348043	2013-12-23	User:Griff88/sandbox

	username	revid	08:38:58 revtime	pagetitle
271	Griff88	587352363	2013-12-23 09:33:19	User:Griff88/sandbox
322	Griff88	598223958	2014-03-05 07:04:43	User:Griff88/sandbox
329	Griff88	610605412	2014-05-29 07:30:19	User:Griff88/sandbox
330	Griff88	610612755	2014-05-29 08:57:23	User:Griff88/sandbox
368	Griff88	613258684	2014-06-17 09:10:17	User:Griff88/sandbox
418	Griff88	620346665	2014-08-08 08:01:05	Wikipedia:VisualEditor/Newsletter
472	Griff88	623962199	2014-09-03 05:14:30	User:Griff88/sandbox
477	Griff88	623994300	2014-09-03 11:09:25	Wikipedia:WikiProject Olympics/Members
584	DilbertReality	563757846	2013-07-11 02:33:21	Wikipedia:Articles for deletion/Chris Chiacchio
661	Laurenwking	536558332	2013-02-04 18:41:28	User:Laurenwking/laurenwkingsandbox
662	Laurenwking	536559529	2013-02-04 18:48:48	User:Laurenwking/laurenwkingsandbox
663	Laurenwking	536562451	2013-02-04 19:07:23	User:Laurenwking/laurenwkingsandbox
670	Laurenwking	539218740	2013-02-20 12:25:29	User:Laurenwking/laurenwkingsandbox
...	...	...	...	...
657513	Ddosguru	620405660	2014-08-08 18:21:42	Wikipedia:Articles for deletion/Botopedia.org

	username	revid	revtime	pagetitle
657514	Ddosguru	620405714	2014-08-08 18:22:08	Wikipedia:Articles for deletion/Log/2014 August 8
657529	Ddosguru	657220490	2015-04-19 18:59:45	Wikipedia:Articles for deletion/IRCCloud
657530	Ddosguru	657220897	2015-04-19 19:03:01	Wikipedia:Articles for deletion/IRCCloud
657531	Ddosguru	657221862	2015-04-19 19:11:34	Wikipedia:Articles for deletion/IRCCloud
657697	Isanaht	619902707	2014-08-05 02:12:59	User:Isanaht/sandbox
657698	Isanaht	619902799	2014-08-05 02:14:17	User:Isanaht/sandbox
657773	Africulture	623454973	2014-08-30 15:03:48	Wikipedia:Changing username/Simple
657774	Africulture	623454973	2014-08-30 15:03:48	Wikipedia:Changing username/Simple
657800	Nothingknewunderthesun	625154022	2014-09-11 22:14:01	User:Nothingknewunderthesun/sandbox
657801	Nothingknewunderthesun	625155247	2014-09-11 22:24:24	User:Nothingknewunderthesun/sandbox
657802	Nothingknewunderthesun	625157768	2014-09-11 22:45:04	Wikipedia:Teahouse/Questions
657803	Nothingknewunderthesun	625157768	2014-09-11 22:45:04	Wikipedia:Teahouse/Questions
657804	Nothingknewunderthesun	625863565	2014-09-16 21:02:41	User:Nothingknewunderthesun/sandbox
657805	Nothingknewunderthesun	625863655	2014-09-16 21:03:33	User:Nothingknewunderthesun/sandbox
657806	Nothingknewunderthesun	625863894	2014-09-16 21:05:43	User:Nothingknewunderthesun/sandbox

	username	revid	revtime	pagetitle
657807	Nothingknewunderthesun	625864880	2014-09-16 21:14:30	User:Nothingknewunderthesun/sandbox
657808	Nothingknewunderthesun	625865326	2014-09-16 21:17:59	User:Nothingknewunderthesun/sandbox
657809	Nothingknewunderthesun	625868880	2014-09-16 21:48:03	User:Nothingknewunderthesun/sandbox
657810	Nothingknewunderthesun	625869984	2014-09-16 21:58:29	User:Nothingknewunderthesun/sandbox
657811	Nothingknewunderthesun	625870858	2014-09-16 22:06:55	User:Nothingknewunderthesun/sandbox
657812	Nothingknewunderthesun	625877618	2014-09-16 23:21:29	User:Nothingknewunderthesun/sandbox
657827	KhalidAliHaji	626880826	2014-09-24 10:41:04	User talk:Sandakelum/Boxes
657828	KhalidAliHaji	626882376	2014-09-24 11:03:17	User:KhalidAliHaji/sandbox
657829	KhalidAliHaji	626883014	2014-09-24 11:11:57	User:KhalidAliHaji/Other/Boxes/Introductior
657830	KhalidAliHaji	626883112	2014-09-24 11:13:17	User:KhalidAliHaji/Other/Boxes/
657831	KhalidAliHaji	626883191	2014-09-24 11:14:19	User:KhalidAliHaji/Other/
657832	KhalidAliHaji	626883293	2014-09-24 11:15:25	User:KhalidAliHaji/Other/
657833	KhalidAliHaji	626883541	2014-09-24 11:18:41	User:KhalidAliHaji/Other/Boxes/Introductior
657857	Dinkjunky	621772362	2014-08-18 13:45:37	User:Dinkjunky/sandbox

93852 rows × 16 columns

In [972]:

```
van_final[ben_final['pagetitle'].str.contains('/')]
```

```
/Users/owner/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:1:
UserWarning: Boolean Series key will be reindexed to match DataFrame index.
  """Entry point for launching an IPython kernel.
```

Out[972]:

	username	revid	revtime	pagetitle	isReverted	revertTime	clue
3	Alex12rocks	535938394	2013-01-31 23:24:08	Pueblo Revolt	True	2013-01-31T23:24:15Z	0
16	RevengeOfBHS	535045160	2013-01-26 21:06:32	Ohio State University	True	2013-01-26T21:06:44Z	0
18	RevengeOfBHS	535045725	2013-01-26 21:09:41	Ohio State University	True	2013-01-26T21:09:49Z	1
26	Sultymcbuttsex	535930942	2013-01-31 22:33:28	Engine	True	2013-01-31T22:33:36Z	0
97	Johnny89detroit	535907433	2013-01-31 19:53:39	Template:Green Bay Packers roster	True	2013-01-31T19:54:04Z	0
100	Johnny89detroit	535908105	2013-01-31 19:57:51	Template:Green Bay Packers staff	True	2013-01-31T19:58:00Z	0
106	Kierankane1	535909587	2013-01-31 20:08:03	Hamilton, South Lanarkshire	True	2013-01-31T20:08:07Z	1
118	TheWikiPikey	535906243	2013-01-31 19:45:29	User talk:Randykitty	True	2013-01-31T19:56:50Z	0
119	TheWikiPikey	535906243	2013-01-31 19:45:29	User talk:Randykitty	True	2013-01-31T19:56:50Z	0
152	SPSpyro	535892970	2013-01-31 18:09:30	Tony Scheffler	True	2013-01-31T18:10:06Z	0
158	Zjslayer	535785124	2013-01-31 01:03:06	List of Minnesota Civil War units	True	2013-01-31T01:03:13Z	0
159	Zjslayer	535785272	2013-01-31	List of Minnesota Civil War units	True	2013-01-31T01:04:20Z	0

	username	revid	revtime	pagetitle	isReverted	revertTime	clue
			01:04:10				
160	Zjslayer	535785585	2013-01-31 01:06:08	List of Minnesota Civil War units	True	2013-01-31T01:07:15Z	0
161	Zjslayer	535863724	2013-01-31 14:30:21	List of Minnesota Civil War units	True	2013-01-31T15:35:58Z	0
176	Slesv	535870716	2013-01-31 15:28:57	User talk:Slesv	False	-	0
186	Ashman741	535857592	2013-01-31 13:32:05	Prejudice	True	2013-01-31T13:50:06Z	0
268	Drloosen2011	535781659	2013-01-31 00:35:00	Coffee	True	2013-01-31T00:35:16Z	1
271	Drloosen2011	535782080	2013-01-31 00:38:33	Big cat	True	2013-01-31T00:38:46Z	0
322	Jezzy68r	535739685	2013-01-30 19:51:27	User talk:Jezzy68r	True	2013-01-30T20:29:06Z	0
329	Wasahobo	534550640	2013-01-23 21:16:55	Canoe	True	2013-01-23T21:28:39Z	0
330	Wasahobo	534551424	2013-01-23 21:21:36	Canoe	True	2013-01-23T21:28:39Z	0
368	Ggg2013	535697462	2013-01-30 15:04:23	Rare species	True	2013-01-30T15:06:16Z	0
418	Jaimenandnick	535591554	2013-01-29 22:55:45	Stopcock	True	2013-01-29T22:55:49Z	1
472	Skateman988	535553377	2013-01-29 18:35:27	Joseph Haydn	True	2013-01-29T18:35:34Z	1
477	Skateman988	535554189	2013-01-29 18:40:47	Joseph Haydn	True	2013-01-29T18:41:06Z	0
584	Delaware2k13	535444568	2013-01-29 02:03:47	China Anne McClain	True	2013-01-29T02:10:51Z	0
661	Katetha	534349763	2013-01-22	Phil Lester	False	-	0

	username	revid	revtime	pagetitle	isReverted	revertTime	clue
662	Katetha	534349810	2013-01-22 16:45:45	Phil Lester	False	-	0
663	Katetha	534349810	2013-01-22 16:45:45	Phil Lester	False	-	0
670	Katetha	534352812	2013-01-22 17:05:54	Phil Lester	True	2013-01-22T17:06:40Z	0
...	...	...	...	...	...	...	...
188324	Bizzenboom1234	521288062	2012-11-04 01:15:37	List of programs broadcast by Family Channel	False	-	0
188325	Bizzenboom1234	521288439	2012-11-04 01:19:00	Motorcity	False	-	0
188326	Bizzenboom1234	521288565	2012-11-04 01:20:10	Tron: Uprising	False	-	0
188359	Bizzenboom1234	523397134	2012-11-16 22:16:34	List of programs broadcast by Nickelodeon	False	-	0
188360	Bizzenboom1234	523397459	2012-11-16 22:18:28	List of programs broadcast by Nickelodeon	False	-	0
188361	Bizzenboom1234	523397459	2012-11-16 22:18:28	List of programs broadcast by Nickelodeon	False	-	0
188362	Bizzenboom1234	524325976	2012-11-22 09:16:09	List of Gravity Falls episodes	True	2012-11-22T14:26:38Z	0
188363	Bizzenboom1234	524325976	2012-11-22 09:16:09	List of Gravity Falls episodes	True	2012-11-22T14:26:38Z	0
188364	Bizzenboom1234	524326242	2012-11-22 09:19:21	List of Gravity Falls episodes	True	2012-11-22T14:26:38Z	0
188365	Bizzenboom1234	524326242	2012-11-22 09:19:21	List of Gravity Falls episodes	True	2012-11-22T14:26:38Z	0
188366	Bizzenboom1234	524326643	2012-11-22 09:23:45	List of Gravity Falls episodes	True	2012-11-22T14:26:38Z	0

	username	revid	revtime	pagetitle	isReverted	revertTime	clue
188367	Bizzenboom1234	524326643	2012-11-22 09:23:45	List of Gravity Falls episodes	True	2012-11-22T14:26:38Z	0
188368	Bizzenboom1234	524326778	2012-11-22 09:25:14	List of Gravity Falls episodes	True	2012-11-22T14:26:38Z	0
188369	Bizzenboom1234	524326778	2012-11-22 09:25:14	List of Gravity Falls episodes	True	2012-11-22T14:26:38Z	0
188370	Bizzenboom1234	524457917	2012-11-23 05:23:12	List of programs broadcast by Nickelodeon	False	-	0
188371	Bizzenboom1234	524457917	2012-11-23 05:23:12	List of programs broadcast by Nickelodeon	False	-	0
188372	Bizzenboom1234	524458467	2012-11-23 05:29:04	List of programs broadcast by Nickelodeon	False	-	0
188373	Bizzenboom1234	524458467	2012-11-23 05:29:04	List of programs broadcast by Nickelodeon	False	-	0
188374	Bizzenboom1234	527449413	2012-12-11 01:21:31	List of Gravity Falls episodes	True	2012-12-11T03:08:07Z	0
188375	Bizzenboom1234	527449413	2012-12-11 01:21:31	List of Gravity Falls episodes	True	2012-12-11T03:08:07Z	0
188377	Bizzenboom1234	529336756	2012-12-22 21:01:15	SpongeBob SquarePants	False	-	0
188403	Bizzenboom1234	547905388	2013-03-31 00:36:14	List of programs broadcast by Disney Channel	False	-	0
188404	Bizzenboom1234	564685687	2013-07-17 18:32:05	List of programs broadcast by Disney Channel	False	-	0
188405	Bizzenboom1234	564685687	2013-07-17 18:32:05	List of programs broadcast by Disney Channel	False	-	0
188406	Bizzenboom1234	564685977	2013-07-17 18:34:32	List of programs broadcast by Nickelodeon	False	-	0
188407	Bizzenboom1234	564685977	2013-07-17 18:34:32	List of programs broadcast by Nickelodeon	False	-	0

	username	revid	revtime	pagetitle	isReverted	revertTime	clue
188448	Bizzenboom1234	591321595	2014-01-18 21:02:33	List of programs broadcast by Nickelodeon	True	2014-01-19T02:27:38Z	0
188488	Klintrin	617665644	2014-07-20 04:37:28	Scholz	True	2014-07-20T04:37:51Z	0
188489	Klintrin	617665831	2014-07-20 04:40:41	User:RandomAct	True	2014-07-20T04:40:50Z	0
188490	Klintrin	617665839	2014-07-20 04:40:48	The Elephant House	True	2014-07-20T04:41:01Z	0

27392 rows × 16 columns

In [973]:

```
ben_final['FirstMeta'] = ben_final.groupby('userid')['pagetitle'].first().str.contains('/')
```

In [974]:

```
len(ben_final['userid'].unique())
```

Out[974]:

```
16496
```

In [975]:

```
van_final['FirstMeta'] = van_final.groupby('userid')['pagetitle'].first().str.contains('/')
```

In [976]:

```
np.sum(van_true)
```

Out[976]:

```
350
```

In [977]:

```
len(van_final['userid'].unique())
```

Out[977]:

```
17015
```

from above analysis, 1533 unique benign users first edit was a metapage, and 350 vandal users first edit was a meta-page. The weighted average ration being 81:19 (benign:vandal). It implies, This implements the feature "fm" in the paper. Which states, if the first edit of the user is a

meta-page, then he is more likely to be a benign user

In [978]:

```
#ben_final = ben_final[0:100000]
#van_final = van_final[0:100000]
```

Taking the Date time stamp into account, we are analysing time difference between the adjacent edits by grouping them according to "pagetitle" and "userid" This will be useful to study the following features:

--> Vandals make faster edits than benign users

--> Vandals spend less time editing a new page

--> Benign users are likely to take longer to edit a new page than a vandal

In [979]:

```
ben_final['diffs'] = ben_final.groupby(['userid'])['revtime'].transform(lambda x: x.diff()) / np.timedelta64(1, 'm')
```

In [980]:

```
ben_final['diffs'] = ben_final['diffs'].fillna(0)
```

In [981]:

```
ben_dummy = ben_final.loc[:,['diffs','userid','pagetitle']]
```

In [982]:

```
ben_dummy.head(20)
```

Out[982]:

	diffs	userid	pagetitle
0	0.000000	18342896	User:Swiftscw
1	1.200000	18342896	User:Swiftscw
2	5.150000	18342896	User:Swiftscw
3	23.383333	18342896	User:Swiftscw/AMAPOLA CABASE
4	7.833333	18342896	Wikipedia:Requests for undeletion
5	0.000000	18342896	Wikipedia:Requests for undeletion
6	45.766667	18342896	Maria Cabase
7	19.933333	18342896	Maria Cabase
8	7.650000	18342896	Maria Cabase

	diffs	userid	pagetitle
9	13.433333	18342896	Maria Cabase
10	52.566667	18342896	Maria Cabase
11	5.366667	18342896	Maria Cabase
12	7.900000	18342896	Talk:Maria Cabase
13	2.433333	18342896	User:Swiftscw
14	12.383333	18342896	Manuel Kabajar Cabase
15	0.666667	18342896	Manuel Kabajar Cabase
16	50.833333	18342896	User talk:Postdlf/Archive23
17	1149.083333	18342896	Talk:Maria Cabase
18	246173.133333	18342896	Wikipedia:WikiProject Biography/Members
19	65.166667	18342896	User talk:Swiftscw

In [983]:

```
ben_dummy.groupby(['userid'])['diffs'].mean()
```

Out[983]:

```
userid
18177826       778.140698
18178056     65067.585185
18178102     13150.735606
18178134     44033.255556
18178244         1.161905
18178257         2.733333
18178679      7556.299206
18178872    196143.520833
18178885      1168.585616
18178917      1035.111905
18178918         1.354167
18179104        83.692424
18179286         3.600000
18179499     52468.865278
18179682       656.505556
18179852         2.466667
18180185    102290.213889
18180236         7.001899
18180345      7517.307738
18180383       355.311667
18180784       236.631013
18180919       240.799621
18181044         0.000000
18181152         0.000000
18181237         2.497727
18181335      4241.415988
18181530     83914.450000
18181773         0.341667
18181786         1.002941
18181935      1599.383333
                 ...
21965844       461.264103
21966099      2772.362179
21966197      3886.234444
```

```
21966297       427.898725
21966358     43413.051667
21967185       478.888889
21967464         2.135897
21968134       142.591667
21968441         0.694444
21969954       229.160000
21971517         0.000000
21971712         5.035185
21973396       736.409211
21973964       660.222222
21974218     11001.712121
21974921       621.662766
21975097         2.090000
21975763       229.081048
21975831        10.455556
21978221       515.329630
21978222       414.757246
21978239      1360.530303
21978274       546.005833
21978316       201.679630
21978776        33.529167
21979000       290.588333
21979213      3347.161111
21979904         7.754167
21980276         1.772222
21982217      2875.983333
Name: diffs, Length: 16496, dtype: float64
```

In [984]:

```
ben_dummy
```

Out[984]:

	diffs	userid	pagetitle
0	0.000000	18342896	User:Swiftscw
1	1.200000	18342896	User:Swiftscw
2	5.150000	18342896	User:Swiftscw
3	23.383333	18342896	User:Swiftscw/AMAPOLA CABASE
4	7.833333	18342896	Wikipedia:Requests for undeletion
5	0.000000	18342896	Wikipedia:Requests for undeletion
6	45.766667	18342896	Maria Cabase
7	19.933333	18342896	Maria Cabase
8	7.650000	18342896	Maria Cabase
9	13.433333	18342896	Maria Cabase
10	52.566667	18342896	Maria Cabase
11	5.366667	18342896	Maria Cabase
12	7.900000	18342896	Talk:Maria Cabase

	diffs	userid	pagetitle
13	2.433333	18342896	User:Swiftscw
14	12.383333	18342896	Manuel Kabajar Cabase
15	0.666667	18342896	Manuel Kabajar Cabase
16	50.833333	18342896	User talk:Postdlf/Archive23
17	1149.083333	18342896	Talk:Maria Cabase
18	246173.133333	18342896	Wikipedia:WikiProject Biography/Members
19	65.166667	18342896	User talk:Swiftscw
20	11600.750000	18342896	Maria Cabase
21	26.150000	18342896	Maria Cabase
22	32616.250000	18342896	Bobby Enriquez
23	2.066667	18342896	Bobby Enriquez
24	2.216667	18342896	Bobby Enriquez
25	0.000000	18359941	User:Azkle
26	982.066667	18359941	Wikipedia:Training/For students/Training feedback
27	5986.750000	18359941	Netflix
28	0.000000	18359941	Netflix
29	65.200000	18359941	Alan Turing
...	...	...	...
657833	3.266667	21807483	User:KhalidAliHaji/Other/Boxes/Introduction
657834	0.550000	21807483	User:KhalidAliHaji
657835	49006.500000	21807483	Athletics at the 2014 Commonwealth Games – Men...
657836	1.616667	21807483	Athletics at the 2014 Commonwealth Games – Men...
657837	328630.416667	21807483	User:KhalidAliHaji
657838	699.716667	21807483	User:KhalidAliHaji
657839	8545.000000	21807483	User:KhalidAliHaji
657840	26312.816667	21807483	User:KhalidAliHaji
657841	3.300000	21807483	User:KhalidAliHaji
657842	0.000000	21940037	User:Kmrishish
657843	1100.683333	21940037	User:Kmrishish
657844	0.000000	21932219	Association of American Universities
657845	0.916667	21932219	Association of American Universities
657846	1.000000	21932219	The Chronicle of Higher Education
657847	0.450000	21932219	The Chronicle of Higher Education
657848	1.900000	21932219	The Chronicle of Higher Education
657849	0.733333	21932219	Carnegie Corporation of New York
657850	10822.483333	21932219	User:Dinkjunky

	diffs	userid	pagetitle
657851	6.716667	21932219	Vera Bartsch
657852	2544.350000	21932219	Joseph John Issa
657853	2.983333	21932219	Joseph John Issa
657854	9.066667	21932219	Joseph John Issa
657855	7052.933333	21932219	Joseph John Issa
657856	4717.950000	21932219	Joseph John Issa
657857	0.016667	21932219	User:Dinkjunky/sandbox
657858	5.283333	21932219	Joseph John Issa
657859	0.000000	21952946	User:Guptav657
657860	5.383333	21952946	User:Guptav657
657861	0.800000	21952946	User:Guptav657
657862	1.000000	21952946	User:Guptav657

657863 rows × 3 columns

checking for beningn users

In [985]:

```
ben_fin = ben_dummy.groupby('userid').nth(1)
```

In [986]:

```
ben_fin.shape
```

Out[986]:

```
(16085, 2)
```

In [987]:

```
ben_fin['EditTime>15'] = ben_fin['diffs']>15
```

There are 5639 benign users who take more than 15 min to edit a new page

checking for vandal users

In [988]:

```
van_final['diffs'] = van_final.groupby(['userid'])['revtime'].transform(lambda x: x.diff()) / np.timedelta64(1, 'm')
```

In [989]:

```
van_final['diffs'] = van_final['diffs'].fillna(0)
```

In [990]:

In [990]:

```
van_dummy = van_final.loc[:,['diffs','userid','pagetitle']]
```

In [991]:

```
van15_fin = van_dummy.groupby('userid').nth(1)
```

In [992]:

```
van15_fin['EditTime>15'] = van15_fin['diffs']>15
```

There are 1951 vandal users who take more than 15min to edit a new page. This is an indication of the behavior for a feature that, benign users take more time to edit a new page than a vandal user

In [999]:

```
ben_fin["stiki_mean"] = ben_scores.groupby('userid')['stiki_score'].mean()
```

In [1000]:

```
ben_fin.describe()
```

Out[1000]:

	diffs	stiki_mean
count	1.608500e+04	16085.000000
mean	9.462651e+03	0.053216
std	5.189957e+04	0.067131
min	0.000000e+00	0.000000
25%	1.616667e+00	0.000000
50%	6.000000e+00	0.032994
75%	4.176667e+01	0.081396
max	1.328026e+06	0.632488

checking for the behavior where, the average of the stikiscore given to user for the edits that has been made, crosses a threshold value

In [1003]:

```
ben_fin['stiki_percent']=ben_fin['stiki_mean']<0.1
```

In [1004]:

```
ben_fin['stiki_percent'].value_counts()
```

Out[1004]:

```
True     13019
False     3066
Name: stiki_percent, dtype: int64
```

In [1001]:

```
van15_fin["stiki_mean"] = van_scores.groupby('userid')['stiki_score'].mean(
)
```

In [1002]:

```
van15_fin.describe()
```

Out[1002]:

	diffs	stiki_mean
count	1.619200e+04	16192.000000
mean	5.641306e+03	0.267262
std	7.712583e+04	0.176928
min	0.000000e+00	0.000000
25%	0.000000e+00	0.139955
50%	1.600000e+00	0.260055
75%	4.900000e+00	0.380814
max	3.450404e+06	0.929759

In [1005]:

```
van15_fin['stiki_percent']=van15_fin['stiki_mean']<0.1
```

In [1006]:

```
van15_fin['stiki_percent'].value_counts()
```

Out[1006]:

```
False    13063
True      3129
Name: stiki_percent, dtype: int64
```

In [1007]:

```
van15_fin.head()
```

Out[1007]:

	diffs	pagetitle	EditTime>15	stiki_mean	stiki_percent
userid					
107645	3.300000	Wikipedia:Community portal	False	0.000000	True
284493	1.133333	Wikipedia:Articles for deletion/Urban75	False	0.001028	True
301475	7.166667	Dick Allen	False	0.000000	True
340857	0.016667	Rolling	False	0.000000	True
345175	3.300000	Surrey Central Station	False	0.000000	True

345175	3.300000	Surrey Central Station	False	0.000000	True
	diffs	pagetitle	EditTime>15	stiki_mean	stiki_percent

In [1028]:

```
ben_final['isReverted'].value_counts()
```

Out[1028]:

```
False    612657
True      45206
Name: isReverted, dtype: int64
```

In [1029]:

```
van_final['isReverted'].value_counts()
```

Out[1029]:

```
True     136499
False     51992
Name: isReverted, dtype: int64
```

In [1010]:

```
ben_fin['FirstMeta'] = ben_final.groupby('userid')['pagetitle'].first().str
.contains('/')
```

In [1011]:

```
van15_fin['FirstMeta'] = van_final.groupby('userid')['pagetitle'].first().s
tr.contains('/')
```

In [1012]:

```
ben_fin.head()
```

Out[1012]:

	diffs	pagetitle	EditTime>15	stiki_mean	stiki_percent	Fi
userid						
18177826	502.500000	User:Shamprabhakar/sandbox	True	0.000000	True	Fa
18178056	78.850000	User:Jack Odanaka	True	0.034222	True	Fa
18178102	9.800000	User:DanielQueiros	False	0.036366	True	Fa
18178134	2.533333	User talk:Arre 9	False	0.000000	True	Fa
18178244	1.683333	Lee Chong Wei	False	0.000000	True	Fa

checking for a behaviour where the number of times the edit is reverted by cluebot is more frequent.

In [1030]:

```
from scipy import stats
ben_fin['reverted_mode'] = ben_final.groupby(['userid']).agg({'isReverted':
lambda x:stats.mode(x)[0]})
```

In [1031]:

```
van15_fin['reverted_mode'] = van_final.groupby(['userid']).agg({'isReverted
': lambda x:stats.mode(x)[0]})
```

In [1032]:

```
ben_fin['reverted_mode'].value_counts()
```

Out[1032]:

```
False    15130
True       955
Name: reverted_mode, dtype: int64
```

In [1033]:

```
van15_fin['reverted_mode'].value_counts()
```

Out[1033]:

```
True     14918
False     1274
Name: reverted_mode, dtype: int64
```

In [1034]:

```
ben_fin.head()
```

Out[1034]:

	diffs	pagetitle	EditTime>15	stiki_mean	stiki_percent	Fi
userid						
18177826	502.500000	User:Shamprabhakar/sandbox	True	0.000000	True	Fa
18178056	78.850000	User:Jack Odanaka	True	0.034222	True	Fa
18178102	9.800000	User:DanielQueiros	False	0.036366	True	Fa
18178134	2.533333	User talk:Arre 9	False	0.000000	True	Fa
18178244	1.683333	Lee Chong Wei	False	0.000000	True	Fa

In [1035]:

```
ben_out = ben_fin.drop(['diffs','pagetitle','stiki_mean','cluebot_mode'],ax
is=1)
```

In [1036]:

```
van_out = van15_fin.drop(['diffs','pagetitle','stiki_mean','cluebot_mode'],
axis=1)
```

In [1037]:

```
ben_out.to_csv('benign_out.csv')
van_out.to_csv('vandal_out.csv')
```

In [1038]:

```
ben_out['type']='Benign'
van_out['type']='Vandal'
```

In [1039]:

```
data_final = ben_out.append(van_out)
```

This is the final data set that we have build based on the analysis of behaviour depicted by the user while edit a page. This will be integrated into the model to make the vandal prediction for test data.

In [1040]:

```
data_final.head()
```

Out[1040]:

	EditTime>15	stiki_percent	FirstMeta	reverted_mode	type
userid					
18177826	True	True	False	True	Benign
18178056	True	True	False	False	Benign
18178102	False	True	False	False	Benign
18178134	False	True	False	False	Benign
18178244	False	True	False	True	Benign

In [1041]:

```
data_final.to_csv('Data_final.csv')
```

In [1043]:

```
data_final.columns
```

Out[1043]:

```
Index(['EditTime>15', 'stiki_percent', 'FirstMeta', 'reverted_mode', 'type'
], dtype='object')
```

In [1045]:

```
X = data_final[['EditTime>15', 'stiki_percent', 'FirstMeta',
'reverted_mode']]
y = data_final['type']
```

From Scikit-Learn Importing train_test_split function

Splitting input and target data into train and test data

In [1046]:

```
from sklearn.model_selection import train_test_split
```

In [1047]:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33,
random_state=42)
```

From Scikit-Learn importing Logistic Regression Function

In [1048]:

```
from sklearn.linear_model import LogisticRegression
```

Building Logistic Regression Model

In [1049]:

```
lr = LogisticRegression()
lr.fit(X_train,y_train)
```

Out[1049]:

```
LogisticRegression(C=1.0, class_weight=None, dual=False,
fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
          penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
          verbose=0, warm_start=False)
```

Predicting the Target Variable using the model build above

In [1050]:

```
ypred = lr.predict(X_test)
```

From Scikit-Learn Importing Confusion_matrix and Accuracy

In [1051]:

```
from sklearn.metrics import confusion_matrix,accuracy_score
```

Comparing Actual and Predicted Target values using Confusion_Matrix

In [1052]:

```
confusion_matrix(y_test,ypred)
```

Out[1052]:

```
array([[5013,  337],
       [ 415, 4887]])
```

Calculating Accuracy

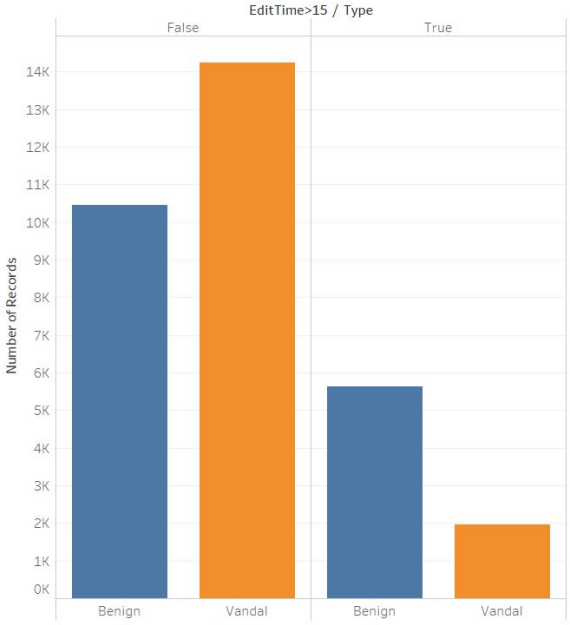
In [1053]:

```
accuracy_score(y_test,ypred)
```

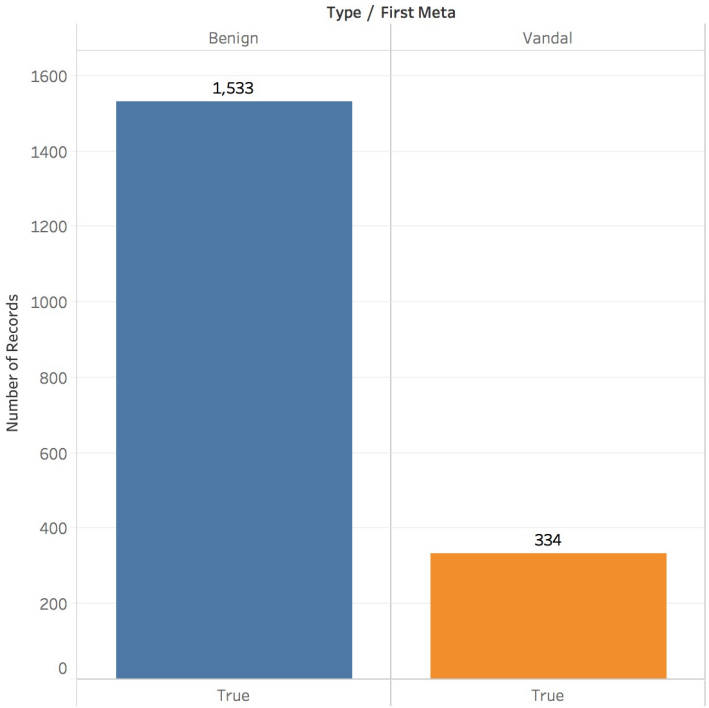
Out[1053]:

0.92940292902741273

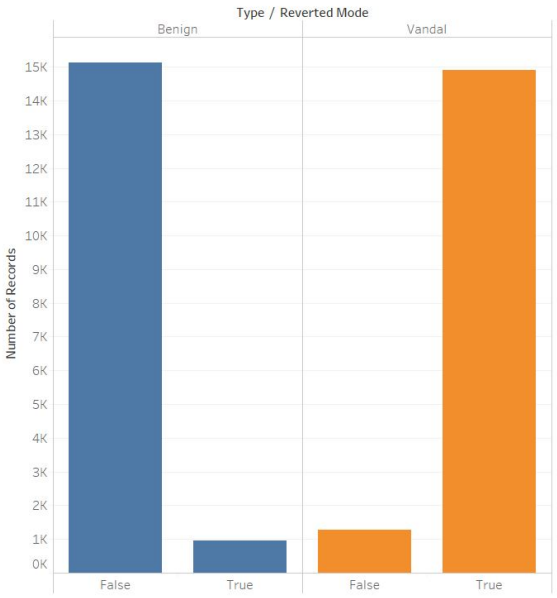
V. Post implementation Analysis of data



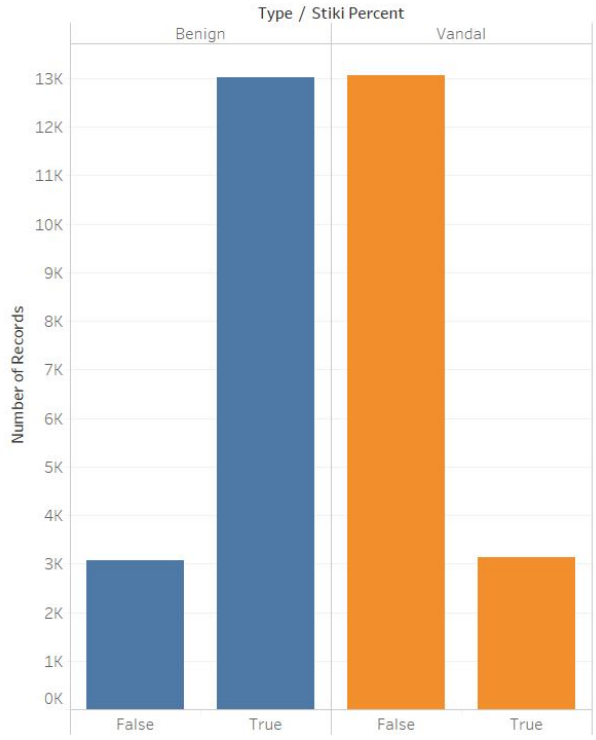
The above figure shows the plot of userid vs Edittime>15. We can clearly see that vandal users take less time to edit a new page.



The above figure shows the probability that the first edit of a Benign user is a Meta page is high.



The above figure shows that the number of times the edit is reverted by cluebot is more frequent with the vandal users.



The above figure shows that the average of stiki score given to the user for the edits that has been made is more for a vandal user.

References

[1] VEWS: A Wikipedia Vandal Early Warning System Srijan Kumar Computer Science Dep. University of
Maryland College Park, 20742 MD, USA srijan@cs.umd.edu Francesca Spezzano UMIACS University of
Maryland College Park, 20742 MD, USA spezzano@umiacs.umd.edu V.S. Subrahmanian Computer
Science Dep. University of Maryland College Park, 20742 MD, USA vs@cs.umd.edu
[2] https://en.wikipedia.org/wiki/Wikipedia:About
[3] https://pypi.python.org/pypi/unicodecsv/0.9.0
[4] http://en.wikipedia.org/wiki/Wikipedia:Vandalism.
[5] P. Neis, M. Goetz, and A. Zipf, “Towards automatic vandalism detection in openstreetmap,” ISPRS
International Journal of Geo-Information, vol. 1, no. 3, pp. 315–332, 2012.
[6] http://en.wikipedia.org/wiki/User:ClueBot_NG. [7]
http://en.wikipedia.org/wiki/Wikipedia:STiki.