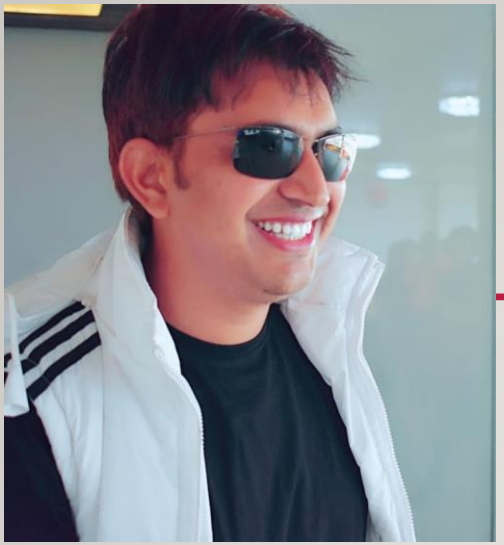
MACHINE LEARNING PROJECT

FAKE NEWS DETECTION USING TWITTER DATA AND NLP

SUBMITTED TO:



PROF. SURESH KUMAR CHOUDHARY

SUBMITTED BY:

ANKIT SINGH [2018MSBDA003] ,
ARKAPRABHA MAJUMDAR [2018MSBDA015]

OBJECTIVES

- Problem statement
- Datasets
- Implementation and technology used
- Tabulate the results
- References

PROBLEM STATEMENT:-

- Characterizing Political Fake News in Twitter – Methods of Identification and Classification

OUR INSPIRATION:-

The recent Google Initiative seminar by the Dept. Of Cultural & Media Studies,CURAJ on the 18th of January has inspired us to take up this topic and try to get rid of the extensive fake news problem in today's media

Central University of Rajasthan

Department of Culture & Media Studies

presents

Google News Initiative Talk-cum-Workshop

on

FAKE NEWS

Sponsored by Google

Speaker: Mr. Ashutosh Jha (Sr. Journalist, ETV)

Date: 18 January 2019

Time: 10:00a.m. - 1:00p.m.

Venue: Seminar Hall, B-2

All are Cordially Invited

DATASETS

LINKS:

https://t.co/AlvByiSrMn [@JebBush - Maybe Donald negotiated a deal with his buddy @HillaryClinton. Continuing this path will put her in the White House.

https://t.co/x2ZimtFxyl [@FullFrontalSamB - Unfortunately Melania copied HER ballot from Michelle so... Donald just voted for Hillary. #ElectionDay

https://t.co/BM3UxA7heR [@BBCTaster - BREAKING NEWS: If you face-swap @realDonaldTrump with @MayorofLondon you get Owen Wilson.

https://t.co/Dh1tWM4FAP [@FoxNews - Report: @HillaryClinton's plan would raise taxes $1.3T/10 years.

THE TWEETS DATA CONSIST OF 1327 rows × 17 columns

We have got a dataset of 1300 tweets which were tagged as fake and we used 60% of our data to train our model and the rest 40% to test the result to reach upto required accuracy

SUMMARY OF DATA USING R:-

Console Terminal

~/

```
>   View(electionday_tweets_clean)
> summary(electionday_tweets_clean)
 is_fake_news    fake_news_category    tweet_id                                  created_at    retweet_count
 Mode :logical   Min.   :1.000      Min.   :2.640e+17   Fri Apr 08 02:50:00 +0000 2016:   1    Min.   : 1002
 FALSE:1191      1st Qu.:1.000      1st Qu.:7.888e+17   Fri Apr 08 23:10:31 +0000 2016:   1    1st Qu.: 1370
 TRUE :136       Median :1.000      Median :7.950e+17   Fri Aug 09 04:41:25 +0000 2013:   1    Median : 2090
                 Mean   :2.118      Mean   :7.788e+17   Fri Aug 19 02:28:45 +0000 2016:   1    Mean   : 3633
                 3rd Qu.:4.000      3rd Qu.:7.960e+17   Fri Aug 21 15:35:18 +0000 2015:   1    3rd Qu.: 3732
                 Max.   :5.000      Max.   :7.961e+17   Fri Aug 26 20:58:13 +0000 2016:   1    Max.   :79092
                 NA's   :1191                           (Other)                       :1321
                                                                                                                                            text
 Absolutely incredible! This is the @realDonaldTrump I know and love. This is the Donald Trump America needs! #maga https://t.co/RKlyI3LB4S  :   2
 '@AmyMek Every Time I see @realDonaldTrump address a crowd I want to start chanting USA, USA, USA! #AmericanPride is Back #Trump2016'       :   1
 '@arcuate: dude that's freaking cool as heck RT @realDonaldTrump wind turbine blades will slice 14 million birds and bats to death in 10 yrs:   1
 '@bigopl: @realDonaldTrump @CNN @oreillyfactor https://t.co/vXiQru6HIE'  Wow, really nice!                                                  :   1
 '@davidshiloach: @realDonaldTrump Go Mr. Trump! Israel is behind you!'                                                                      :   1
 '@HillaryClinton is the continuity candidate. If you want change in America you vote for @realDonaldTrump. https://t.co/A0sWt9jaUu           :   1
 (Other)                                                                                                                                     :1320
       user_screen_name user_verified  user_friends_count user_followers_count user_favourites_count
 FoxNews        :  48    Mode :logical   Min.   :     0.0   Min.   :       16     Min.   :     0
 DanScavino     :  46    FALSE:539       1st Qu.:   254.5   1st Qu.:    31862     1st Qu.:   215
 realDonaldTrump:  37    TRUE :788       Median :   779.0   Median :   194830     Median :  1563
 TeamTrump      :  30                    Mean   :  8983.6   Mean   :  3509883     Mean   : 11724
 DonaldJTrumpJr :  25                    3rd Qu.:  3326.0   3rd Qu.:   926522     3rd Qu.:  7190
 mike_pence     :  25                    Max.   :578811.0   Max.   :93936531     Max.   :335194
 (Other)        :1116
                                                                              tweet_source geo_coordinates
 <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>   :521   Min.   :0.0000
 <a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>                   :461   1st Qu.:0.0000
 <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>: 93   Median :0.0000
 <a href="https://about.twitter.com/products/tweetdeck" rel="nofollow">TweetDeck</a> : 62   Mean   :0.8357
 <a href="https://studio.twitter.com" rel="nofollow">Media Studio</a>                : 39   3rd Qu.:1.0000
 <a href="https://ads.twitter.com" rel="nofollow">Twitter Ads</a>                    : 33   Max.   :9.0000
 (Other)                                                                              :118
  num_hashtags      num_mentions        num_urls         num_media
 Min.   :0.0000   Min.   : 1.000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.: 1.000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median : 2.000   Median :0.0000   Median :0.0000
 Mean   :0.8357   Mean   : 1.946   Mean   :0.4115   Mean   :0.2321
 3rd Qu.:1.0000   3rd Qu.: 2.000   3rd Qu.:1.0000   3rd Qu.:0.0000
 Max.   :9.0000   Max.   :11.000   Max.   :2.0000   Max.   :1.0000

> |
```

In [8]:
```
import pandas as pd
a=pd.read_csv("electionday_tweets.csv")
```

In [9]:
```
a.head()
```

Out[9]:

	is_fake_news	fake_news_category	tweet_id	created_at	retweet_count	text	user_screen_name	user_verified	user_friends_cou
0	False	NaN	264033382076407808	Thu Nov 01 15:57:18 +0000 2012	4698	@realDonaldTrump you are full of shit!	RalphGilles	True	78
1	False	NaN	265895586660757505	Tue Nov 06 19:17:02 +0000 2012	9646	@realDonaldTrump you're fucking retarded	TimmyWait	False	83
2	False	NaN	265895723445411841	Tue Nov 06 19:17:35 +0000 2012	1823	@realDonaldTrump You are the stupidest man on ...	mattcale52	False	118
3	False	NaN	265896172726661120	Tue Nov 06 19:19:22 +0000 2012	1168	@realDonaldTrump I am continually amazed and t...	MichaelWHill	False	160
4	False	NaN	266042962650226688	Wed Nov 07 05:02:39 +0000 2012	1979	Hey @realDonaldTrump You Mad Bro?	ThePresObama	False	13

SUMMARY OF THE DATA USING PYTHON (JUPYTER NOTEBOOK):-

```
In [12]: a.describe()
```

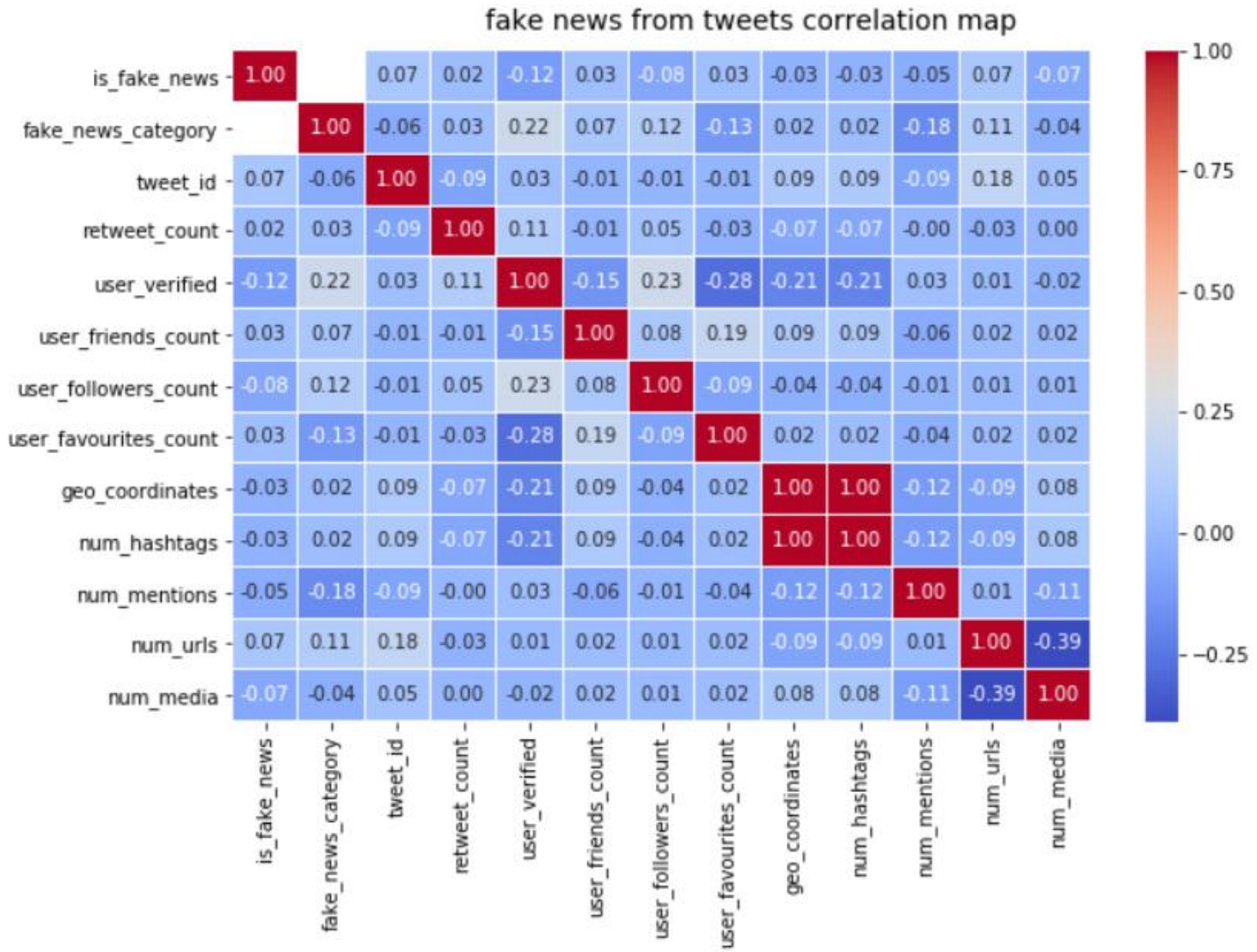
Out[12]:

	fake_news_category	tweet_id	retweet_count	user_friends_count	user_followers_count	user_favourites_count	geo_coordinates	num_hashtags
count	136.000000	1.327000e+03	1327.000000	1327.000000	1.327000e+03	1327.000000	1327.000000	1327.000000
mean	2.117647	7.788172e+17	3633.334589	8983.629992	3.509883e+06	11723.996232	0.835720	0.835720
std	1.633260	5.602454e+16	5305.272308	33033.454741	1.181786e+07	28691.640919	1.180596	1.180596
min	1.000000	2.640334e+17	1002.000000	0.000000	1.600000e+01	0.000000	0.000000	0.000000
25%	1.000000	7.888427e+17	1370.000000	254.500000	3.186200e+04	215.000000	0.000000	0.000000
50%	1.000000	7.950101e+17	2090.000000	779.000000	1.948300e+05	1563.000000	0.000000	0.000000
75%	4.000000	7.959941e+17	3732.000000	3326.000000	9.265215e+05	7190.000000	1.000000	1.000000
max	5.000000	7.961288e+17	79092.000000	578811.000000	9.393653e+07	335194.000000	9.000000	9.000000

CORRELATION HEATMAP USING SEABORN:-

In [22]:

```
f, ax = plt.subplots(figsize=(10, 6))
corr = a.corr()
hm = sns.heatmap(round(corr,2), annot=True, ax=ax, cmap="coolwarm",fmt='.2f',
                 linewidths=.05)
f.subplots_adjust(top=0.93)
t= f.suptitle('fake news from tweets correlation map', fontsize=14)
```



Implementation and technology used

Detecting fake news is not an easy task since there can be many definition of fake news . But there is some extent to which it can be possible using some machine learning models.
We can also use a lot of NLP(natural language processing) technology to make the computer understand the news as there is a lot of text involved.

The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python programming language.

matplotlib

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

python™
Natural Language Analyses with NLTK

```
1
2  #-------------------------------------------------------------------------
3  # Include Libraries
4  #-------------------------------------------------------------------------
5
6  import pandas as pd
7  from sklearn.model_selection import train_test_split
8  import sklearn
9  from sklearn.feature_extraction.text import CountVectorizer
10 from sklearn.feature_extraction.text import TfidfVectorizer
11 from sklearn.naive_bayes import MultinomialNB
12 from sklearn import metrics
13 from pandas_ml import ConfusionMatrix
14 from matplotlib import pyplot as plt
15 from sklearn.linear_model import PassiveAggressiveClassifier
16 from sklearn.feature_extraction.text import HashingVectorizer
17 import itertools
18 import numpy as np
```

TABULATE THE RESULT

Finally at the end of our project ,we are estimating the likelihood of our results to be about 90% accurate to detect the news if it was a real one or fake one

REFERENCES:-

- https://arxiv.org/abs/1712.05999v1
- https://seaborn.pydata.org/tutorial.html
- matplotlib.org/users/pyplot_tutorial.html
- https://www.nltk.org/book/

THANK YOU