Analysis of News Category

Team Project - Report

submitted to

MANIPAL
ACADEMY of HIGHER EDUCATION
(Deemed to be University under Section 3 of the UGC Act, 1956)
INSPIRED BY LIFE

by

Noopur Agrawal
Reg No. 181046002
BDA II

Ashwathguru S
Reg No. 181046039
BDA II

Under the guidance of
Arockiaraj Sir

4th March 2019

Index

Introduction

This project deals with the exploratory data analysis of News Category dataset. Data is extracted from www.huffingtonpost.com, a popular news channel website in US. This dataset contains news for the six years 2012-2016. Around 2 Lakh records were fetched on daily for the six years which contain different columns:

- Category : Total 41 distinct category which includes Politics, Health, Education other major topics.
- Author Name : Around 27000 distinct authors contributed towards news for 6 years
- Headline
- Short Description
- Date, Month, Year
- Day of the week

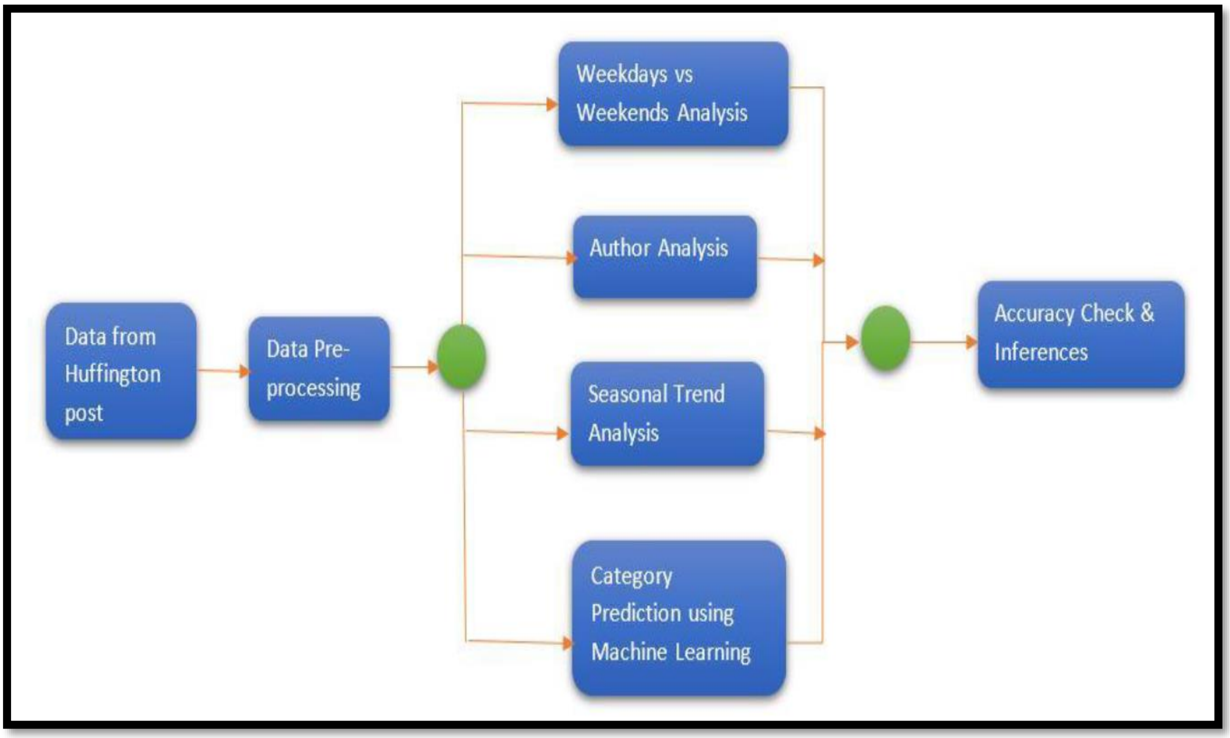
Strategy for data analysis:

- News published based on author
- News published based on category
- Popularity of news on yearly basis
- Popularity of news on weekday and weekend basis

Data Classification Methods

- Logistics Regression
- SVM Classifier
- Random Forest Classifier

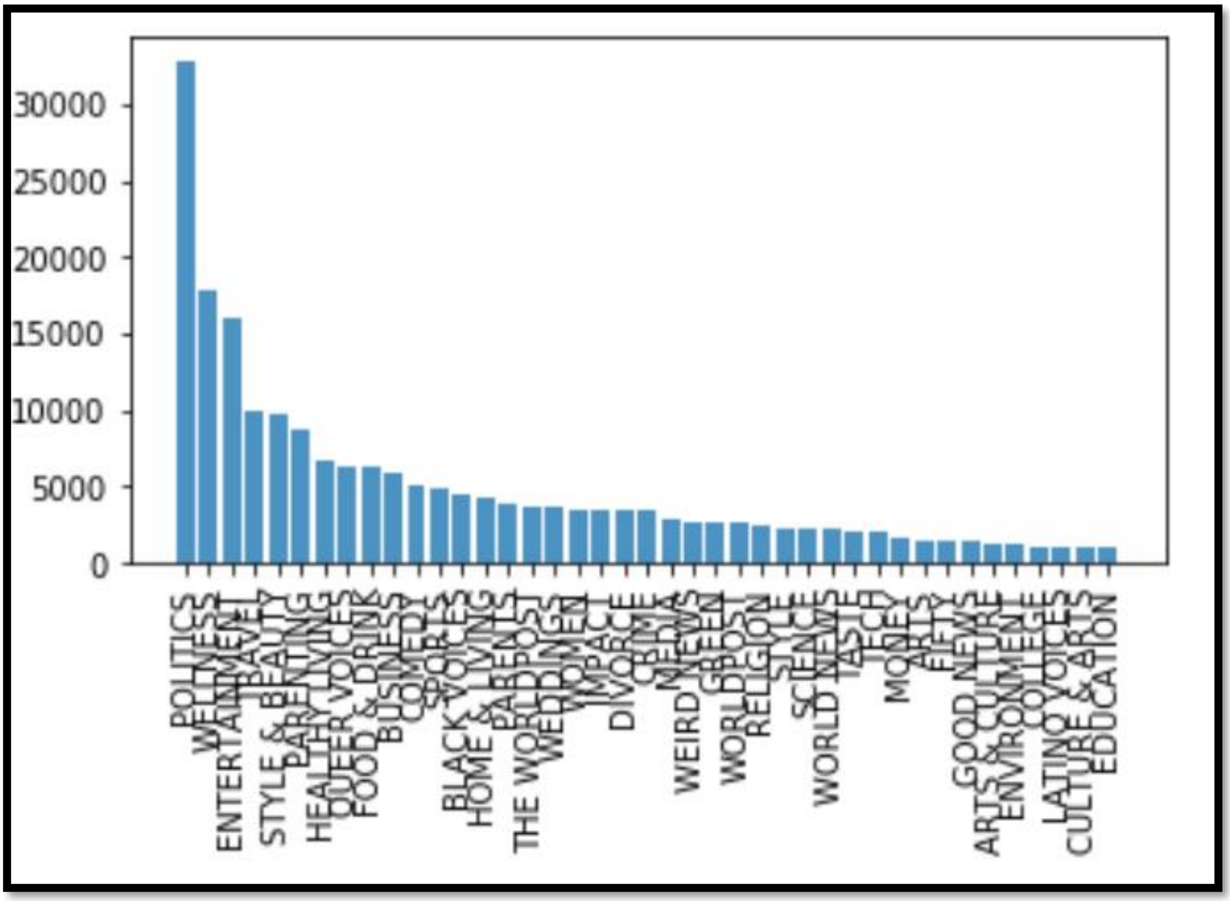
Steps Involved in News Classification & Analysis



Exploratory Data Analysis

1. **News Published category-wise**

   This analysis was done to analyse the popularity of category throughout the period of 6 years. Hence the frequency of news published in each of 41 category is plotted. The screenshot of result is attached below.



The top 5 category are:

- Politics
- Wellness
- Entertainment
- Travel
- Style and beauty

2. **Data Slicer:**

   It will slice the data based on start data and end date. Helps us to do analysis for a particular period.

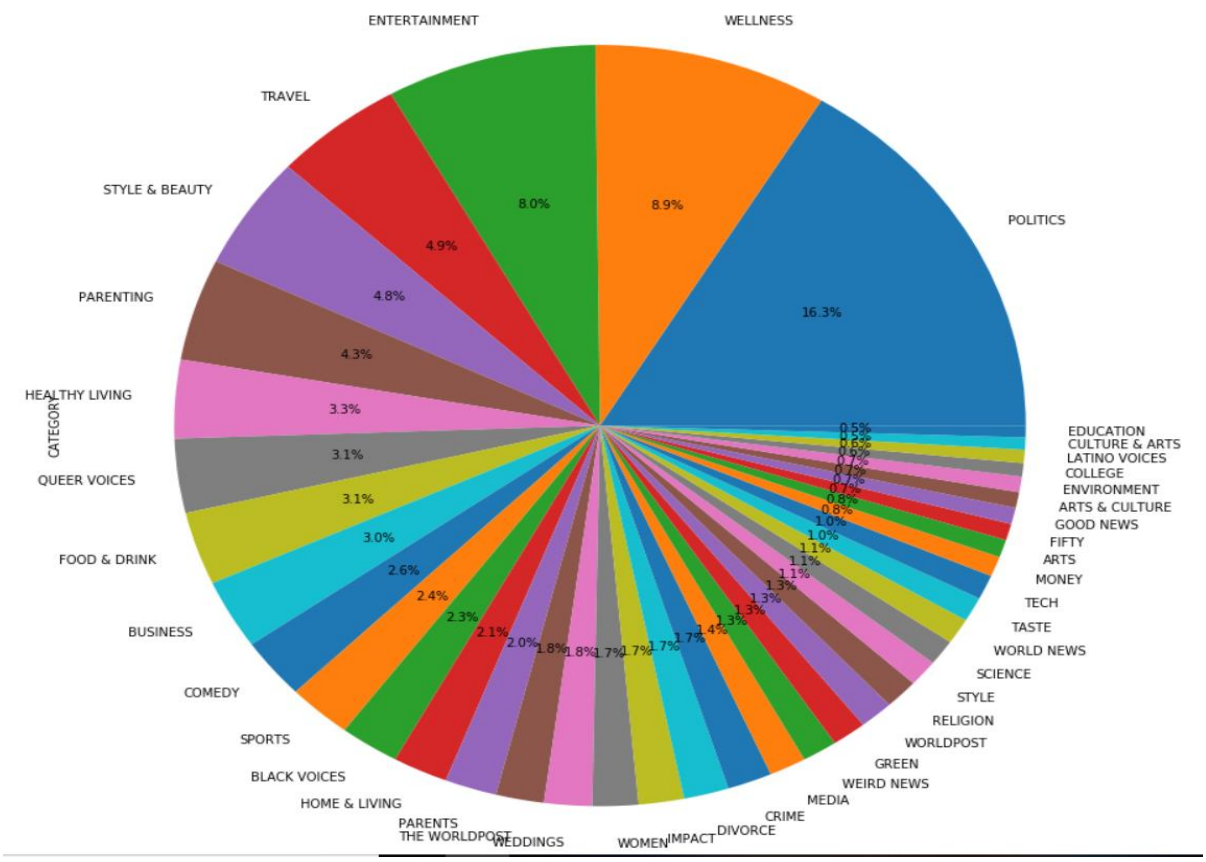
3. **Date to Date Analysis**

   After slicing, we analysed the data on the particular slice based on the frequency of news published in each category. Number of authors who were more active in the particular period and the active year between those period. Screenshot of result has been attached below for more clear explanation.

```
The active year/s in the given range : [2016, 2017]

The categories published : ['BLACK VOICES', 'POLITICS', 'COLLEGE', 'SCIENCE', 'ENTERTAINMENT', 'WORLD NEWS', 'IMPACT',
'HEALTHY LIVING', 'SPORTS', 'BUSINESS', 'TASTE', 'TRAVEL', 'QUEER VOICES', 'WORLDPOST', 'STYLE', 'LATINO VOICES', 'FIF
CATION', 'ARTS', 'RELIGION', 'MEDIA', 'COMEDY', 'TECH', 'CRIME', 'WEIRD NEWS', 'GREEN', 'GOOD NEWS', 'PARENTS', 'ARTS
E', 'THE WORLDPOST']

The number of active authors : 6110
{'BLACK VOICES': 1187, 'POLITICS': 11713, 'COLLEGE': 117, 'SCIENCE': 200, 'ENTERTAINMENT': 3685, 'WORLD NEWS': 404, 'I
02, 'WOMEN': 956, 'HEALTHY LIVING': 1458, 'SPORTS': 970, 'BUSINESS': 702, 'TASTE': 546, 'TRAVEL': 403, 'QUEER VOICES':
ORLDPOST': 151, 'STYLE': 474, 'LATINO VOICES': 372, 'FIFTY': 99, 'EDUCATION': 173, 'ARTS': 52, 'RELIGION': 483, 'MEDIA
'COMEDY': 1332, 'TECH': 277, 'CRIME': 722, 'WEIRD NEWS': 739, 'GREEN': 550, 'GOOD NEWS': 328, 'PARENTS': 1167, 'ARTS &
E': 836, 'THE WORLDPOST': 2575}

Authors who wrote more than 200 posts: ['Julia Brucculieri', 'Curtis M. Wong', 'Jenna Amatulli', 'Mary Papenfuss', 'An
ld', 'Steven Hoffer', 'Sara Boboltz', 'Caroline Bologna', 'Bill Bradley', 'Lee Moran', 'Cristian Farias', 'Matthew Jac
na Golgowski', 'Igor Bobic', 'Carly Ledbetter', 'Marina Fang', 'James Michael Nichols', 'Rebecca Shapiro', 'Taylor Pit
auren Weber', 'None', 'Ed Mazza', 'Hilary Hanson', 'Sam Levine', 'Carla Herreria', 'Paige Lavender', 'Jamie Feldman',
ye', 'Daniel Marans', 'Ron Dicker', 'Cole Delbyck']
```

4. Percentage of news distribution in each category

Out of all the 41 distinct categories, 16% of news published in politics, 8.9 % in wellness and 8 % in Entertainment. Rest others are less than 5% of total. The screenshot of percentage distribution in each category is shown below.



5. Top Author Analysis

This analysis aimed to find the top most author who contributed towards news publishing. The result is shown below. Lee Moran is the most active author.

Lee Moran	2423
Ron Dicker	1913
Reuters, Reuters	1562
Ed Mazza	1322
Cole Delbyck	1140
Andy McDonald	1068
Julia Brucculieri	1059
Carly Ledbetter	1054
Curtis M. Wong	1020

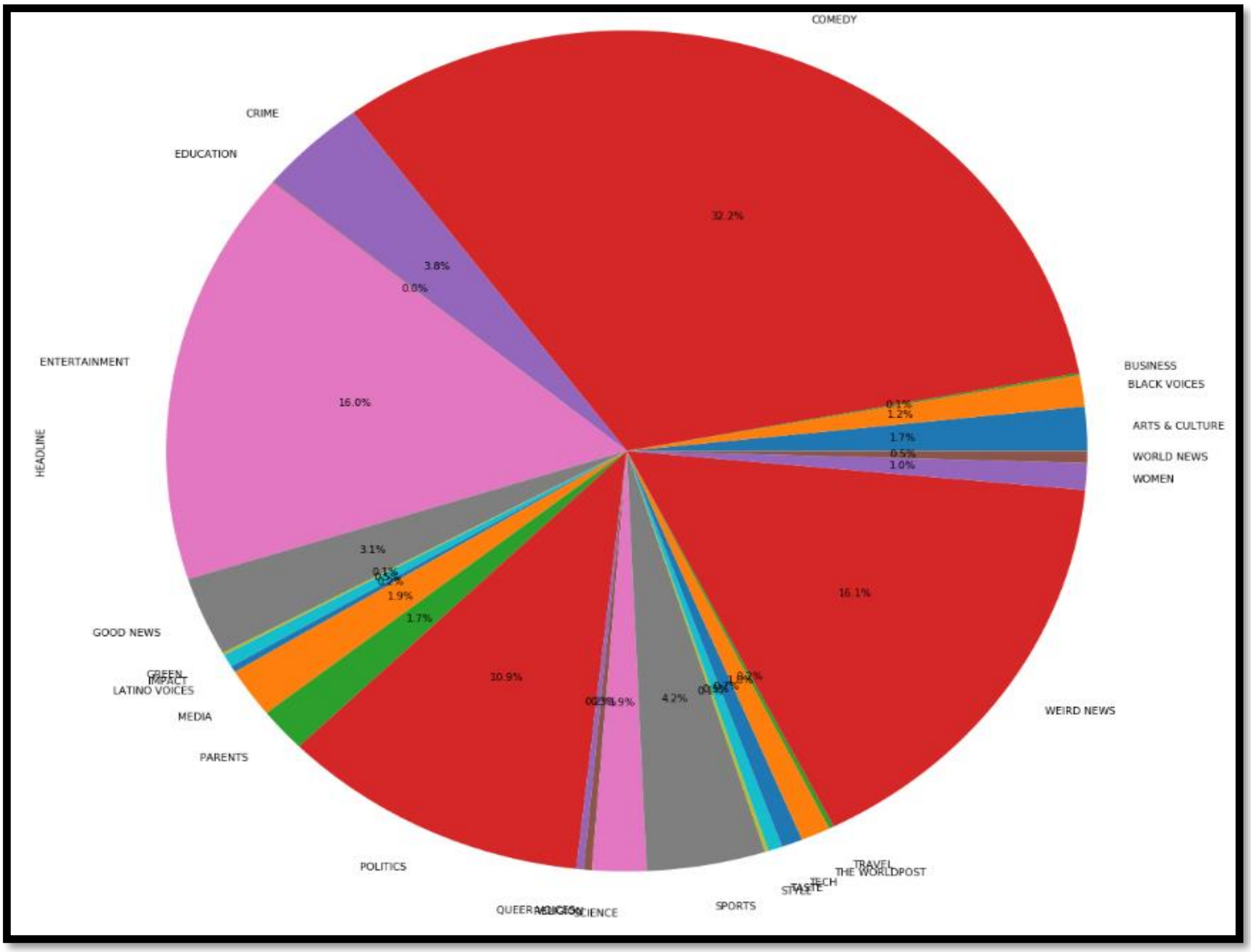
Now we want to see in which category he is publishing more news. The analysis of news category published by Lee Moran is shown below.

CATEGORY	
ARTS & CULTURE	41
BLACK VOICES	29
BUSINESS	2
COMEDY	779
CRIME	92
EDUCATION	1
ENTERTAINMENT	387
GOOD NEWS	74
GREEN	2
IMPACT	12
LATINO VOICES	6
MEDIA	46
PARENTS	41
POLITICS	263
QUEER VOICES	6
RELIGION	7
SCIENCE	46
SPORTS	101
STYLE	3
TASTE	12
TECH	18
THE WORLDPOST	25
TRAVEL	4
WEIRD NEWS	390
WOMEN	25
WORLD NEWS	11

From the above analysed, we concluded that Lee Moran, the top-most author published most of the news in Comedy category, then weird news, Entertainment and Politics.

News we want to see the percentage of news distribution in each category published by Lee Moran.

The below pie chart shows the Lee Moran published 32 % of news in Comedy category, around 16 % in weird news and entertainment each. 10% in politics. Rest other category is below 10 %.

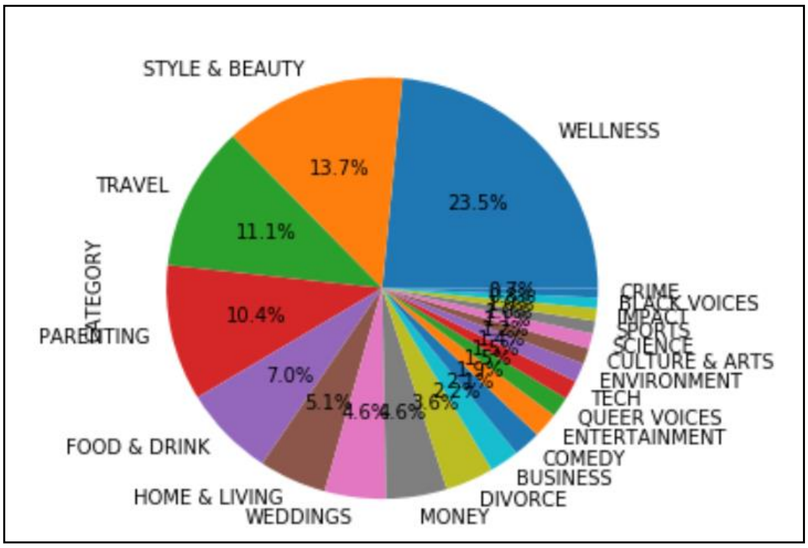


6. Yearly analysis on news authors and categories

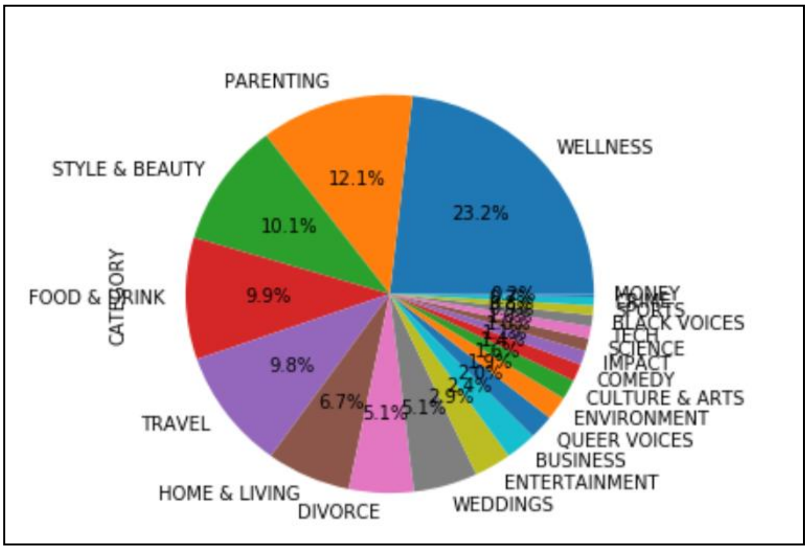
The main aim of this analysis was to find the most active author and the most popular category in each year from 2012 to 2018.

The result of analysis is shown below.

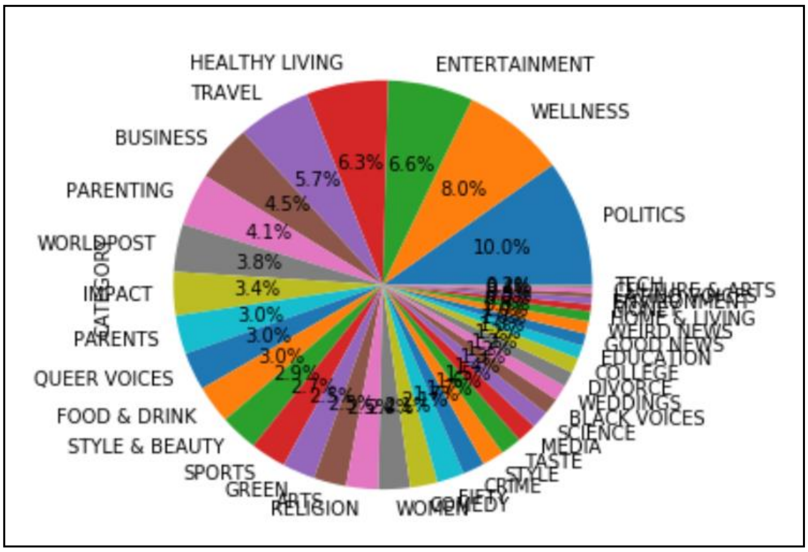
```
Top 5 authors in 2012
 None                5048
Reuters, Reuters     414
Michelle Manetti     269
Rebecca Adams        203
Michelle Persad      197
Amy Marturana        182
Name: AUTHOR, dtype: int64
```



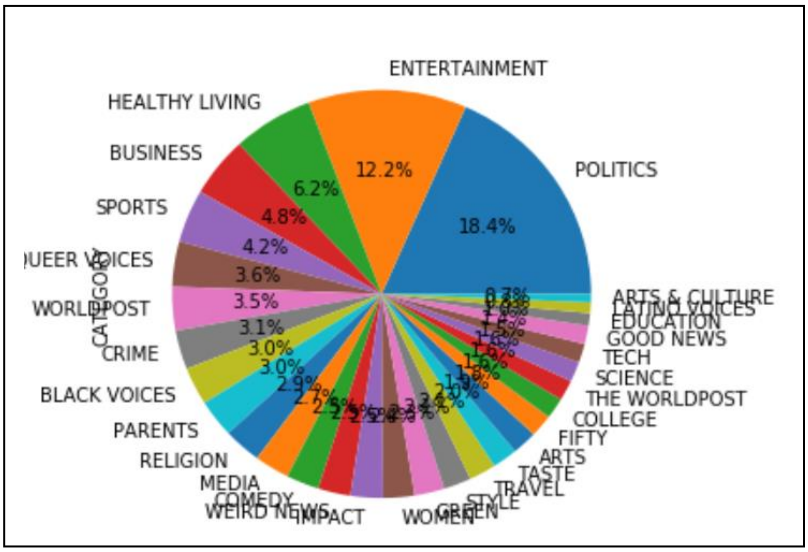

```
Top 5 authors in 2013
 None                11246
Reuters, Reuters       681
Michelle Manetti       555
Rebecca Adams          342
Dana Oliver            317
Michelle Persad        295
Name: AUTHOR, dtype: int64
```



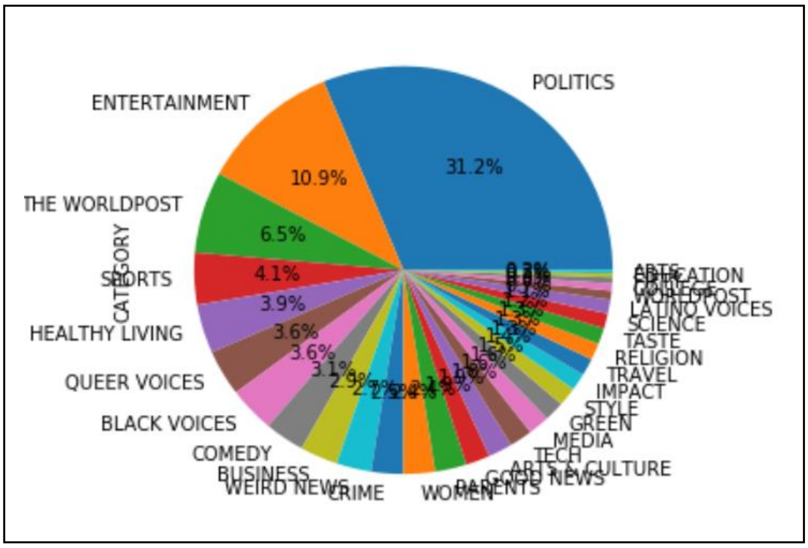
```
Top 5 authors in 2014
 None                  4970
Dana Oliver             201
Chris Greenberg         181
Jamie Feldman           173
JamesMichael Nichols    173
Priscilla Frank         171
Name: AUTHOR, dtype: int64
```



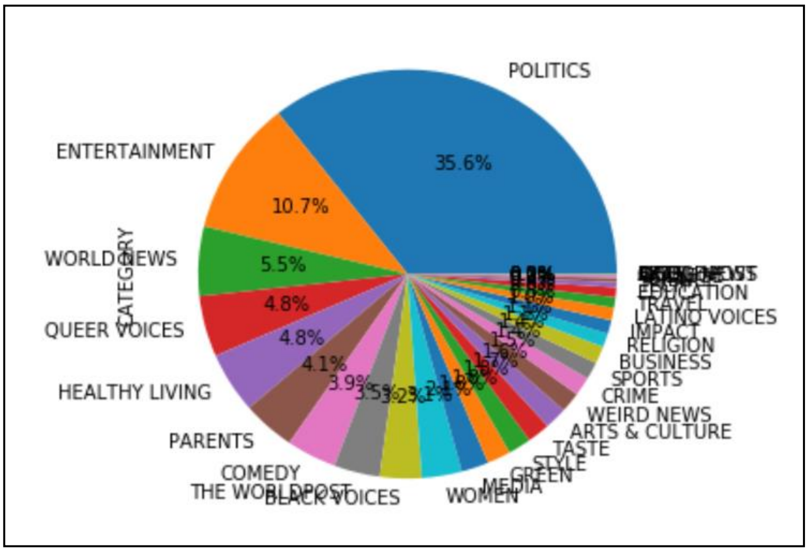

```
Top 5 authors in 2015
 None                4968
Bill Bradley          336
Lily Karlin           321
Julia Bruccullieri    318
Ron Dicker            274
E. Oliver Whitney     256
Name: AUTHOR, dtype: int64
```



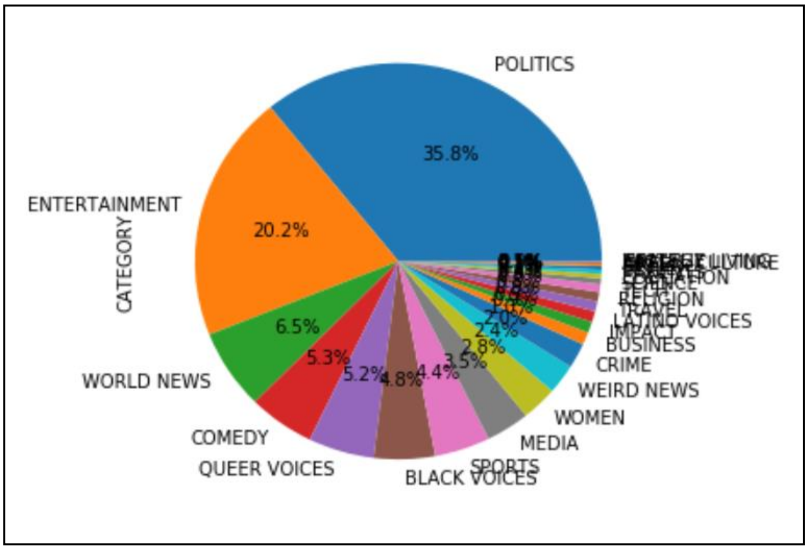
```
Top 5 authors in 2016
 None                4946
Lee Moran             811
Cole Delbyck          546
Julia Bruccul ieri    536
Ron Dicker            535
Carly Ledbetter       440
Name: AUTHOR, dtype: int64
```




```
Top 5 authors in 2017
 None               1172
Lee Moran            867
Mary Papenfuss       603
Ron Dicker           495
Ed Mazza             381
Caroline Bologna     337
Name: AUTHOR, dtype: int64
```



```
Top 5 authors in 2018
 Lee Moran          610
Ed Mazza           402
Ron Dicker         375
Mary Papenfuss     301
Jenna Amatulli     266
Curtis M. Wong     243
Name: AUTHOR, dtype: int64
```



7. **Weekdays and Weekends Analysis**

This analysis was done to find the authors who are the most active and least active during weekdays. Weekdays is taken from Monday to Thursday and weekends is taken from Friday to Sunday.

The result of analysis on author is shown below.

```
Top 5 authors on Weekdays
 None                 15208
Ron Dicker            1060
Ed Mazza               796
Lee Moran              787
Carly Ledbetter        657
Andy McDonald          574
Name: AUTHOR, dtype: int64

Least popular 5 authors on Weekdays
 Barry Wilner & Tom Canavan, AP
Cooper Koch, Contributor\r\nHusband to Todd, Dad to Claire & Mason, Founder of Cooper Smit...
Peyton Price and Stephanie Sprenger, Contributors
Carine Fabius, ContributorAuthor, art dealer, museum curator and temporary body art pioneer
Aaron Anson, Contributor\r\nInspirational self-help speaker and author, 'Mind Your Own Lif...
Name: AUTHOR, dtype: int64
```

```
Top 5 authors on Weekends
 None                  16223
Lee Moran              1339
Reuters, Reuters        848
Mary Papenfuss          650
Ron Dicker              496
Bill Bradley            480
Name: AUTHOR, dtype: int64

Least popular 5 authors on Weekends
 Martin J. Bernstein, Contributor\r\nWriter, hiker and nature enthusiast
Tom Schraeder, Contributor\r\nMusician/Blogger
Carla Herreria and Ryan J. Reilly
Taina Bien-Aime, ContributorExecutive Director, Coalition Against Trafficking in Women (CATW)
George Clooney, Contributor
Name: AUTHOR, dtype: int64
```

Now the same analysis is done on the popular category during weekdays and weekends. Result is shown below.

```
Top 5 categories on Weekdays
 POLITICS              15019
WELLNESS               9413
ENTERTAINMENT          6405
STYLE & BEAUTY         5146
TRAVEL                 5071
PARENTING              4792
Name: CATEGORY, dtype: int64

Least 5 categories on Weekdays
 COLLEGE               553
FIFTY                 512
EDUCATION             469
ENVIRONMENT            74
CULTURE & ARTS         42
Name: CATEGORY, dtype: int64
```

```
Top 5 categories on Weekends
 POLITICS          12533
ENTERTAINMENT      7562
WELLNESS           5303
QUEER VOICES       3319
BUSINESS           3201
TRAVEL             3098
Name: CATEGORY, dtype: int64

Least 5 categories on Weekends
 COLLEGE            410
EDUCATION          403
LATINO VOICES      363
MONEY              362
ARTS & CULTURE     336
Name: CATEGORY, dtype: int64
```

News Classification

The news classification system aimed to build a prediction system which will automatically classify the news headline into its respective categories.

Steps followed in Classification is as follows:

Train test Split

Our first task is to first split input data into training and test data. For this we have used 80% data as training data and rest 20% as test data.

Next we will convert training data into vector of features then give as input to the model to learn. The extraction and normalization of features from training data is called as Feature Scaling.

For this we use Count vector and TF-IDF Vector. The detailed working of Count Vector and TF-IDF vector is explained below:

Feature Scaling

We used count vector and TF-IDF vector for feature scaling. Its main work is to extract features from text data and convert in into numeric data.

Count Vector:

It converts raw texts into bag of words on the basis of frequency of occurrence of those words. It count the total number of unique elements in the training data and arrange it in the form of array.

The total number of rows corresponds to the total rows of training data while the columns refers to the each unique word in training data arranged in alphabetical order. In python sklearn package provides the Count Vector library.

Example:

```
messages = ["Hey hey hey lets go get lunch today",
            "Did you go home?",
            "Hey!!! I need a favor"]
```

Corresponding count vector of "messages".

	did	favour	get	go	hey	home	lets	lunch	need	today	you
0	0	0	1	1	3	0	1	1	0	1	0
1	1	0	0	1	0	1	0	0	0	0	1
2	0	1	0	0	1	0	0	0	1	0	0

In the first instance of training data, we can see that the word "Hey" occurred three times, Hence in the vector form its value is written as 3. Similarly the value in vector corresponds to the total number of occurrence of the word in the training data.

Logistics Regression Classifier

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for given set of features(or inputs), X.

We can also say that the target variable is categorical. Based on the number of categories, Logistic regression can be classified as:

1. **Binomial**: target variable can have only 2 possible types: “0” or “1” which may represent “win” vs “loss”, “pass” vs “fail”, “dead” vs “alive”, etc.

2. **Multinomial**: target variable can have 3 or more possible types which are not ordered(i.e. types have no quantitative significance) like “disease A” vs “disease B” vs “disease C”.

3. **Ordina**l: It deals with target variables with ordered categories. For example, a test score can be categorized as: “very poor”, “poor”, “good”, “very good”. Here, each category can be given a score like 0, 1, 2, 3.

In our cases, we have used Multinomial Logistics regression model, since we have more that 3 distinct categories to predict.

The accuracy we got by logistics regression is 62.98%

Random Forest Classifier

Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.

The accuracy we got by Random Forest Classifier is 51.56%

SVM Classifier

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well (look at the below snapshot). Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line).

The accuracy we got by SVM is 58.63%

Conclusion

The following inferences can be drawn from the above analysis process.

- Out of all the 2 lakh records present, Politics is the most popular category and education is the least popular.
- Out of all 27,993 authors, only 259 authors have contributed 60% of the total news published in six years.
- During 2012 and 2013, Wellness, Style & Beauty and Parenting was the most popular category. But since 2014 Politics dominated over others. This could be due to the upcoming presidential election in 2016.
- Ron Dicker and Lee Moran are the most active authors during weekdays as well as in weekends.
- Politics, Entertainment, Wellness and Travel is the top 5 categories during weekdays and weekends.

- Education and Culture & Arts is the least popular category during weekdays and weekends.
- Logistics Regression is giving more accuracy as compared to SVM and Random Forest classifier. This could be due to the fact that Logistics regression is more suitable for categorical dataset.

Challenges and Future scope

- The analysis can be made further better if we have more information about demographical region and the number of reads based on area wise.
- In some of the records, author name, categories were missing. This was replaced by 'None'. The above analysis is done by ignoring these missing records.
- Further analysis can be improved by adding tags to each articles. This tags will help to better analyse the words used and their frequency based on region.

References

- https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- https://towardsdatascience.com/understanding-logistic-regression-9b02c2aec102