# CS 5593 - Data Mining

# Fall 2019

# Fake News Detection

Link to Demo: https://www.youtube.com/watch?v=jWxRp3Dt5ZQ

Madhumitha Pachapalayam Sivasalapathy – madhups1992@ou.edu

Siva Rama Krishna Ganta – Shivasrk1234@ou.edu

Mamidi Sai Kiran Reddy – Sai.kiran.reddy.mamidi-1@ou.edu

Chanukya Lakamsani – Chanukyalakamsani@ou.edu

**Abstract**

As we live in a data world fake news is emerging on the internet day by day. Social media is becoming a major source of information for everyone. People intentionally or unintentionally write or share false information on social media to mislead others. There is a lot of unverified news on the internet which could cause consequences in people's lives. For example, the stock price of a product can significantly impact if a rumor about a company abruptly hits on social media. This will have substantial impact on individuals and society. In today's world, almost everyone has access to the internet, we can get any information within a second's, there are tons of resources out there to help us. But all the resources are not trustworthy, uncertain they might mislead individuals. Therefore, one major challenge is to identify this fake news based on the topic of interest is by suggesting the user whether the search result has conspiracy content.

**Introduction:**

The invest of social media services facilitated people in searching and sharing information at an unprecedented level. When Pew Research Center began tracking social media adoption in 2005, just 5% of American adults used at least one of these platforms. By 2011 that share had risen to half of all Americans, and today 72% of the public uses some type of social media. Since people have freedom of expression, they can express their opinion on social media. For instance, during 2016 US presidential election, candidates electively used Tweeter to send their messages and express their opinions directly to their supporters. Hillary Clinton's Tweeter, for example, has reached 16 million followers, with her most popular tweet received more than 600,000 re-tweets 2 and one million likes.

**Research Problem:**

Fake news is a significant problem in our connected world. Although misinformation and propaganda have been around for ages, fake news is now becoming a real threat, noticeably due to the ease of creating, spreading and consuming the content online. What makes fake news a hard problem is identifying, tracking and controlling unreliable content. Individuals tend to believe anything that comes up online.

**2 Related Work:**

There exist several online fact-checking applications, like Hoaxy [Luca, 2016], which gather credibility scores of news from several online fact-checking websites, such as Snopes.com, PolitiFact.com, and FactCheck.org. Recent research shows that fake news spread fast through social media networks (for instance, Hurricane Sandy in 2012 [Farida, 2012] and the Boston Marathon blasts in 2013 [3]), and this can adversely amplify public anxiety. For instance, in 2013, $130 billion in "stock value was wiped out" abruptly after a rumor of Barack Obama being injured by an explosion at the White House. Thus, early identification of fake news, before they find their way in online fact-checking repositories, potentially using solely their content, is imperative.

In hybrid deep model for fake news detection [Yan, 2017] research, they have work on uncovering the fake news articles present in social media website like Twitter. Twitter everyone will tweet about someone else or some other news articles. This could be retweeted. In this research they are predicting the news articles that are fake and are retweeted more often by the same group of users. They are considering the users who have been Spreading the fake news more often beyond the threshold with the original news are taken as fake news spreaders. They concentrated more on the social media behavior for the fake news.

Trends in the diffusion of misinformation on social media by [hunt 2019]. They were comparing the news articles that were published by various news websites with the Facebook and Twitter news updates. They found that there were 570 fake news websites and 10,250 fake news stories on Facebook and Twitter. By, analyzing the different website they have categorized few

websites that are mostly publishing fake news. Hunt also analyzed the trend of Twitter and Facebook sharing percentage of fake news. They found that fake news engagement through tweeter is increasing year over year where as for Facebook it went to the peek in the mind of 2017 and starts decreasing. Their more focus is on engagement each fake news resulted in the social media.

## 3 The proposed work

Our objective for this project is to develop a predictive model using classifier on data set containing fake and real news. Using this model, we have developed full stack web application to predict the YouTube fake/real results for the searched topic.

### 3.1 Data set and Source

We have collected the two data sets from Kaggle (https://www.kaggle.com/mrisdal/fake-news).
First dataset contains fake news observations and second data set contains real news observations. We are merging the two data sets to get our final data set. The size of the data set is 54 MB. It contains 28,711 observations. Out of them 15,712 observations belong to real news category and 12,999 observations belong to fake news category. Totally data set contains 20 attributes like the news website, name of the author and the date of the post.

We have chosen only 2 variables from this original dataset for this classification, Since the other attributes are not very useful. The other variables can be added later to add some more complexity and enhance the features.

Below are the columns used to create the dataset that have been in used in this project

- Column 1: Statement (News title).
- Column 2: Label (Label class contains: Real, Fake)

The dataset used for this project were in csv format and can be found in repo.

| | title | label |
|---|---|---|
| 0 | Muslims BUSTED: They Stole Millions In Govât... | fake |
| 1 | Re: Why Did Attorney General Loretta Lynch Ple... | fake |
| 2 | BREAKING: Weiner Cooperating With FBI On Hilla... | fake |
| 3 | PIN DROP SPEECH BY FATHER OF DAUGHTER Kidnappe... | fake |
| 4 | FANTASTIC! TRUMP'S 7 POINT PLAN To Reform Heal... | fake |

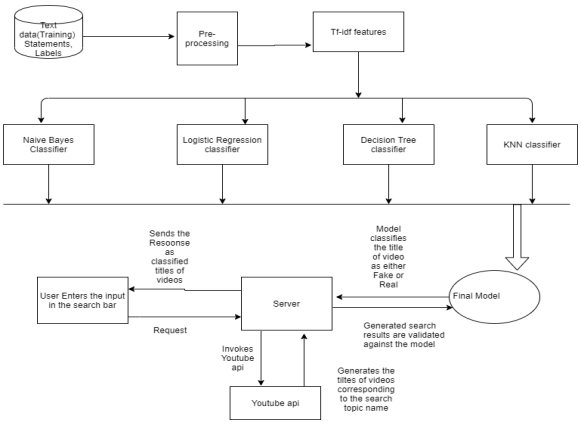
**Feature engineering:**
In this project we have performed feature extraction and selection methods from sci-kit learn python libraries. For feature selection, we have used term frequency- Inverse Document Frequency like tf-idf weighting. we have also used word2vec and POS tagging to extract the features, though POS tagging and word2vec has not been used at this point in the project.

A sample of the data set is shown below:

| | abuse | accuse | act | allen | benedict | bishop | brother | call | catholic | child | ... | time | to | ufo | vatican | victim | work | world |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0 | 0.0 | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | ... | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 |
| 1 | 0.0 | 0.0 | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | ... | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 |
| 2 | 0.0 | 0.0 | 0.000000 | 0.546292 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | ... | 0.000000 | 0.000000 | 0.591344 | 0.0 | 0.0 | 0.0 | 0.0 |
| 3 | 0.0 | 0.0 | 0.000000 | 0.000000 | 0.632033 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | ... | 0.000000 | 0.413638 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 |
| 4 | 0.0 | 0.0 | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | ... | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

### 3.2 Architecture

The application has a GUI where in a user can enter the search topic name in the search bar. When user clicks on the search button. Request is sent to the server. In the back end, a classifier model is built from the data set using logistic regression. Also, in the back end, our program invokes the YouTube api. It generates the videos titles and URLs corresponding to the search topic name. Generated search results are validated against the model. Model classifies whether the video is fake or real. Server sends the classified videos to the front end. The classified videos with titles are displayed on the GUI of the application.




**Resources:**

Front End (GUI): Bootstrap, HTML5, CSS3, JavaScript

Programming language: Python

IDE: Pycharm

Framework: Flask for integrating front end and back end

### 3.3 Data Preprocessing

Generally, the performance of a text classification model is highly dependent on the words in a corpus and the features created from those words.

We have analyzed the data set and performed some data cleaning techniques to improve the performance of the application.

Following preprocessing techniques are applied on our application :

Tokenizing into words: This refers to splitting a text string into a list of 'tokens', where each token is a word.

Eg: s1='I have a car' will be converted to ['I', 'have', 'a, 'car'].

Converting to Lowercase: This step is performed because capitalization does not make a difference in the semantic importance of the word.

Eg. 'Travel' and 'travel' should be treated as the same. It will also reduce the feature dimensionality, instead of taking 'Travel' and 'travel' as different features. It considers only one feature

Removing non-alphabetical words and 'Stop words': Stop words and other noisy elements are removed since they increase feature dimensionality but do not usually help to differentiate between documents.

Eg: is, the

Lemmatization: Applying Lemmatization which converts each word to its root form, turning different words into a single representation. It reduces the feature dimensionality.

Eg: played and playing, lemmatizer converts the two words into their root form, play and play. So we consider only one feature instead of two.

Removing Extra spaces and punctuations: All the special characters

(eg: .,:-#), single characters(eg: a, I) and extra spaces are removed

### 3.4 Algorithm

Here we are using classification as our data mining task. We have implemented Naive-bayes, Logistic Regression, KNN and Decision tree classifiers for predicting the fake you tube results. Each of the classifiers are trained with the extracted features. Once fitting the model, we compared the results like f1 score and analyzed the confusion matrix results. Among all the classifiers, 2 best performing models were selected as candidate models for predicting fake results. Between these two models, we have chosen Logistic Regression as our final classifier model, it is used for fake news detection with the probability of truth. In Addition to this, we have also extracted the top 250 features from our term-frequency tfidf vectorizer to see what words most are and important in each of the classes. We have also used Precision-Recall and learning curves to see how training and test set performs when we increase the amount of data in our classifiers.

#### 3.4.1 KNN(K-Nearest Neighbor) Algorithm:

K-Nearest Neighbor is a simple classification algorithm that finds the classification of categories based on going over the whole data set to find its k nearest Neighbor. It is non-parametric, and no assumptions need to be met.

In K-NN uses distance measure to find the nearest neighbor. These are some of the different distance

functions that we can choose from: Euclidean Distance, Hamming Distance, Manhattan Distance, Minkowski Distance. Here, we used Euclidean distance.
A distance metric or distance function is a real-valued function d, such that for any coordinates x, y, and z:

1. $d(x,y) \geq 0$,and $d(x,y) = 0$ if and only if $x = y$

2. $d(x,y) = d(y,x)$

3. $d(x,z) \leq d(x,y) + d(y,z)$

Property 1 : To make sure that distance should be nonnegative, and the only way for distance to be zero is for the same duplicates (eg: Text1: “This is cat.” Text2: “This is cat.” In this case the distance will be zero).

Property 2 : Make sure it is commutative, for example, the distance from New York to Los Angeles is the same as the distance from Los Angeles to New York. Finally. [Kousar , 2016]

**Steps for the Algorithm:**

Defining Function to calculate distance. [First Euclidian & cosine distance]

For each item in the trainset: Calculate the distance between each item with the new data point. Sort the training set based on shortest distance in ascending order Take the top k training set records. Count the records for each label and the label that fall in top k sets. The highest count on labels would be the closest label for a given data point.

**Hyper-parameter:**

The KNN normally was not giving us the best result. We tried various K values and 40 seems give us better results. It has curse of dimensionality to a greater effect. Since we have much higher dimension.

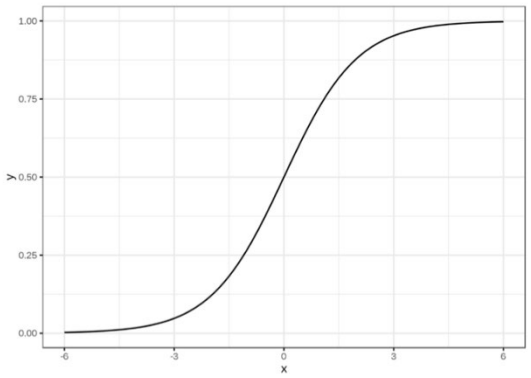
**3.4.2 Logistic Regression:**

Linear regression models fail to work accurately for classification problems, the reason being the reason being the linear model does not output probabilities but it treats the classes as numbers (0 and 1) and fits the best hyperplane (for a single feature, it is a line) that minimizes the distances between the points and the hyperplane. So, it simply interpolates between the points, and you cannot interpret it as probabilities.

Logistic regression models the probabilities for classification problems with binary outcomes based on the threshold. It’s an extension of linear regression model for classification problems.

A model for classification is logistic regression. Instead of fitting a straight line or hyperplane, the logistic regression model uses the logit function to convert the output of a linear equation between 0 and 1 [Indra,2016]. We set the threshold to be 0.5 to compare this model with other models’ performance. If the output is less than 0.5 it is predicted as real news and if it is greater than 0.5 it is predicted as fake news. The logistic function is defined as:

$$\text{logistic}(\eta)=\frac{1}{1+\exp(-\eta)}$$




For classification, we prefer probabilities between 0 and 1, so we wrap the right side of the equation into the logistic function. This forces the output to assume only values between 0 and 1.

In the linear regression model, the relationship between the outcome and the features is modelled by the linear equation. [Sandra, 1993]

For classification, we prefer probabilities between 0 and 1, so we wrap the right side of the equation into the logistic function. This forces the output to assume only values between 0 and 1.

Binary Logistic Regression (We are using this algorithm for the dataset since we have to classify as fake news(1) or real(0).)

### 3.4.3 Classification and Regression Tree (CART) Algorithm:

Decision Trees CART uses Gini index to select the its attribute. It has the Capability to handle both numerical and categorical variables. It uses GINI index measure to see how well a given attribute separates training samples for the given target class (it is a measure of misclassification). The Gini index is given as:

$$Gini(D) = 1 - \sum_{j=1}^{n} p_j^2$$

We implemented CART from Loh's paper [Loh, 2011] by measuring of node impurity based on the distribution of the observed Y values in the node. Node impurity is calculated using a cost function that is minimized to choose split points with mean squared error across all training samples. In our implementation of CART, we used Gini Index to compute the impurity of the node. Less impurity could give us 0.0 and higher impurity could lead us to 0.5. Node is split by exhaustively searching over all X and S for the split $\{X \in S\}$ that minimizes the total impurity of its two child nodes. The process is applied recursively on the data in each child node. Pseudocode for tree construction is shown below.

Initialize the root node.

For each X, find the set S that minimizes the sum of the node impurities in the two child nodes and choose the split $\{X^* \in S^*\}$ that gives the minimum overall X and S.

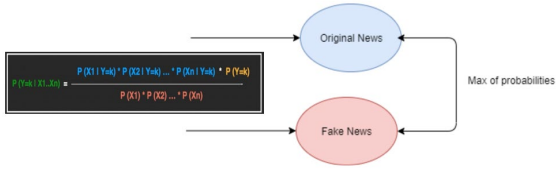
Stop splitting if a termination criterion is reached. Otherwise, go back to step 2.

We included two parameters for termination criterion. The first criterion is maximum tree depth, which is the maximum number of nodes from the root node of the tree. Tree split is terminated when maximum tree depth is reached. Deeper trees are more complex and are more likely to over-fit the training data. Another criterion is minimum node size, which is the minimum number of training patterns that a given node is responsible for. We stop splitting tree when it is below minimum node size.

### 3.4.4 Naïve Bayes classification Algorithm:

Naïve Bayes is one of the simple classification algorithms which make use of the Bayes theorem on probabilities. Although it is simple it often outperforms the More sophisticated classification modeling methods. It works well when the dimensionality of input is high. As in our dataset the dimensionality is high, we believe it should perform well [Mitchell T.M., 2007].

Given the input what is the probability for the output to be original news. And what is the probability for the output to be a fake news. After getting the probabilities for both select the probability which is higher. The one with the higher probability will be our predicted output label. Since the probabilities are calculated for each feature.




### 3.5 Performance Evaluation and Development Choice

Since our dataset is almost balanced We used confusion matrix, accuracy scores majorly also we have compared the precision, recall and F1 scores to evaluate the performance of the four classification algorithms on our data set.

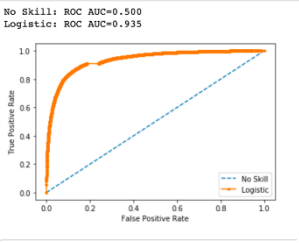
| Model | Accuracy | Precision | Recall | F-1 Score |
|---|---|---|---|---|
| Logistic Regression | 78.5% | 79.2% | 72.9% | 76.04% |
| Naive Bayes | 76.4% | 84.12% | 61.2% | 70.85% |
| Decision Tree | 72.2 % | 72.1% | 71.99% | 71.99% |
| KNN | 48.4% | 71.2% | 52.07% | 36.17% |

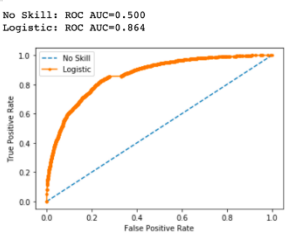
Performance of the Logistic Regression algorithm performs like the naïve Bayes, as expected naïve Bayes worked well for higher dimensions reason each feature are independent of one another so different types of fake content are given a probabilities and help predicting more precisely. KNN, as expected was not able to handle the higher dimensions because the Euclidean distance doesn't mean so much if it must deal with more distant variables.

We also wanted to compare the area under the ROC curve [Sokolova, 2006], Cross validation and below are the results for test set. From the results below we did not see a major change it was similar to the above. Still logistic model and naïve Bayes were performing almost similar. Comparing the training and test set from the different models we could see the model is not over fitted for all the models except KNN because of the dimension.

ROC Curve:

Logistic Regression:



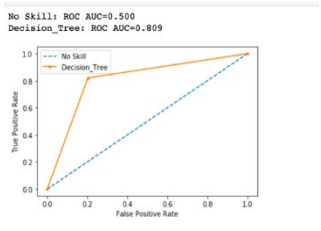



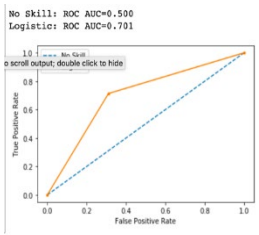

Training set Test set

10-fold Cross-Validation Results:

```
array([0.79317549, 0.78690808, 0.79937304, 0.79832811, 0.78091257,
       0.78439568, 0.78927203, 0.79867642, 0.79380007, 0.79790941])
```

Decision Tree:



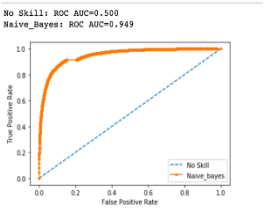



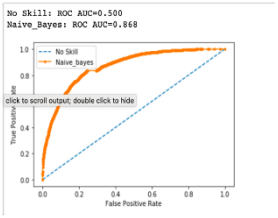

Training set Test set

10 fold CV test results:

```
array([0.72632312, 0.73642061, 0.73423894, 0.72727273, 0.72866597,
       0.72135145, 0.73423894, 0.73632881, 0.72831766, 0.73101045])
```

Naïve Bayes:



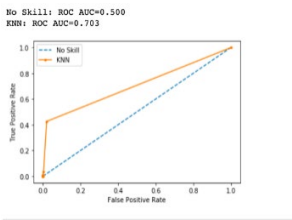



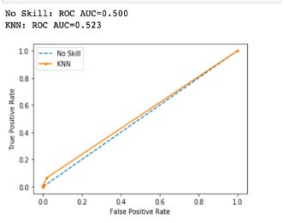

Training set Test set

10-fold CV test results:

```
array([0.76810585, 0.7663649 , 0.77429467, 0.76628352, 0.75374434,
       0.76384535, 0.78091257, 0.78300244, 0.76837339, 0.76480836])
```

KNN:




Training set Test set

10-Fold CV test results:

```
array([0.45264624, 0.45264624, 0.4528039 , 0.4528039 , 0.4528039 ,
       0.4528039 , 0.4528039 , 0.4528039 , 0.4528039 , 0.45261324])
```

**McNemar's Test (Hypothesis test)**:

From the above results we see that models naive Bayes and logistic regression performed equally. we wanted to confirm whether both are statistically the same or different.

Through McNemar's test we can test for models that are expensive to train [Cornelia, 2015], which suits large deep learning models.

Procedure:

Summarization of result.

| | Logistic | Decision | NB |
|---|---|---|---|
| 0 | YES | YES | YES |
| 1 | YES | YES | YES |
| 2 | YES | NO | YES |
| 3 | YES | YES | YES |
| 4 | YES | YES | YES |
| 5 | YES | YES | YES |
| 6 | NO | YES | NO |
| 7 | YES | YES | YES |
| 8 | YES | YES | YES |
| 9 | NO | YES | YES |

A_contingency table is a tabulation or count of two categorical variables. The below figure is an example of contingency table for Logistic, NaiveBayes and Decision Tree classifier. The contingency table lies on the fact that the classifier is trained on the same data.

| | Classifier2 Correct, | Classifier2 Incorrect |
|---|---|---|
| Classifier1 Correct | Yes/Yes | Yes/No |
| Classifier1 Incorrect | No/Yes | No/No |

fig : contingency table

The contingency table obtained for our result is given below

```
[[5047, 586], [433, 1112]]
```

McNemar's test statistic : (Yes/No - No/Yes)^2 / (Yes/No + No/Yes)

The test statistic has a Chi-Squared distribution with 1 degree of freedom.

Significant value of p can be interpreted as

- **p > alpha**: fail to reject H0, no difference in the disagreement (e.g. treatment had no effect).
- **p <= alpha**: reject H0, significant difference in the disagreement (e.g. treatment had an effect).

We can summarize this as follows:

> Fail to Reject the Null Hypothesis: Classifiers have a similar proportion of errors on the test set.
>
> Reject Null Hypothesis: Classifiers have a different proportion of errors on the test set.

for alpha = .05 we get p value equal to 1.9202665775050555e-06. Since p values is very less we can reject the null hypothesis and conclude that both classifiers are statistically different.

Since there is not much difference between the result of Naïve bayes and logistic classifiers we went with logistic regression classifier for the final application.

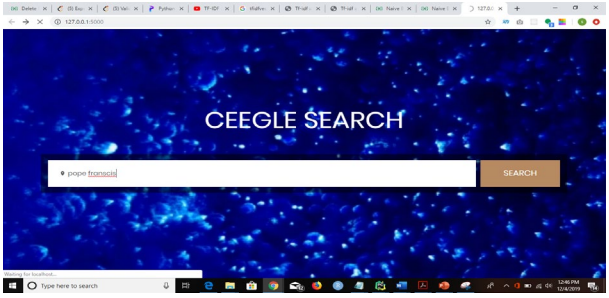
### 3.6 How to use the application

We made the user interface similar to google search engine[for youtube search].

Input: Something you want to search.

The application would retrieve the search results and predict the fake news from the results and highlight the fake news and others as real. You could click on the links to go to web page from the results.

**Screen shots:**




This is were you will enter the input for search.

This is the output screen.

## 4 Conclusion and Future work

The social media news is increasing rapidly, and more and more people consume news from social media and google instead of traditional news media like newspaper or news channel. How every non- traditional medium like social media or google/ youtube search results spread fake news which has strong negative impact on individual users and broader society.

In this project, we reviewed existing fake news detection approaches from a data mining perspective, including feature extraction and model construction.

The fake news takes various forms like text, images, videos, audios. We have considered detecting fake news using text. The possible future work could be more focused on predicting the fake news from images.

## 5 References

[Kousar,2016] A. Kousar Nikhath, K.Subrahmanyam, R.Vasavi, " Building a K-Nearest Neighbor Classifier for Text Categorization", International Journal of Computer Science and Information Technologies, Vol. 7 (1) , 2016, pages 254-256

[Luca, 2016] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. "Hoaxy: A platform for tracking online misinformation", In Proceedings of the 25th International Conference Companion on World Wide Web 2016, pages 745–750.

[Yan, 2017] Natali Ruchansky and Sungyong Seo and Yan Liu, "A Hybrid Deep Model for Fake News" , in ACM conference on Information and Knowledge, 2017, page 797-806.

[Hunt, 2019] Hunt Allcott, Matthew Gentzkow, Chuan Yu, "Trends in the Diffusion of Misinformation on Social Media", NBER working paper, January 2019.

[Loh, 2011], "Classification and regression trees", Wiley Interdiscip. Rev. Data Min. Knowl. Discov., 2011, vol. 1, no. 1, page. 14–23.

[Indra, 2016] Indra, Liza Wikarsa, BCS, MComp, Rinaldo Turang, SKom, MKom, "Using Logistic Regression Method to Classify Tweets into the Selected Topics", ICACSIS, 2016.

[Sandra, 1993] Sandra W. Pyke, Peter M. Sheridan, "Logistic Regression Analysis of graduate student retention", Canadian journal of higher education , 1993, page [49-54].

[Farida, 2012] Farida Vis Burgess, Jean and Axel Bruns, "Hurricane sandy: the most tweeted pictures. the guardian data blog. Available online, http://www.guardian.co.uk/ news/datablog/gallery/2012/nov/06/hurricane-sandy-tweeted-pictures", November 2012.

[Palatucci, 2007] Palatucci M., Mitchell T.M. "Classification in Very High Dimensional Problems with Handfuls of Examples". Knowledge Discovery in Databases, 2007 Lecture Notes in Computer Science, vol 4702.

[Cornelia, 2015] Cornelia M. Borkhoff, patric R. Johnston, Derek Stephens, Eshety Atenafu, "The special case of the 2 X 2 table: asymptotic unconditional McNemar test can be used to estimate sample size even for analysis based on GEE", Journal of clinical Epidemiology, 2015.

[Sokolova, 2006] Sokolova M., Japkowicz N., Szpakowicz S. "Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation", 2006, Advances in Artificial Intelligence. AI 2006. Lecture Notes in Computer Science, vol 4304.