Journal of Advanced Zoology

ISSN: 0253-7214
Volume 44 Issue S-6 Year 2023 Page 971:977

Fake News Identification for Web Scrapped Data

Yashwanth M[1], Laxmi B Rananavare[2]

[1,2]Department of Computer Science, REVA University, India.
Email: yashwanth.m139@gmail.com[1] , laxmib.rananavare@reva.edu.in[2]

**Corresponding author's E-mail: yashwanth.m139@gmail.com*

Article History

Received: 06 June 2023
Revised: 05 Sept 2023
Accepted: 30 Nov 2023

CC License
CC-BY-NC-SA 4.0

Abstract

Majority of the people get affected with misleading stories spread through different posts on social media and forward them assuming that it is a fact. Nowadays, Social media is used as a weapon to create havoc in the society by spreading fake news. Such havoc can be controlled by using machine-learning algorithms. Various methods of machine learning and deep learning techniques are used to identify false stories. There is a need for identification and controlling of fake news posts that have increased in alarming rate. Here we use Passive-Aggressive Classifier for fake news identification. Two datasets, Kaggle fake news dataset and as well as dynamically web scrapped dataset from politifact.com website. We achieved 88.66% accuracy using Passive Aggressive Classifier.

Keywords: *Fake News Detection, TF-IDF, Classification, Passive Aggressive Classifier, Machine Learning, Web Scrapping*

1. Introduction

Fake or false information is defined as a misleading information or a lie that is deliberately spread to confuse or promote prejudice. Satirical stories are published primarily for the purpose of amusing and annoying the audience and they can be dangerous if they are broadcasted on social media. Many false stories for various financial and political reasons have been surfaced on social media to gain individual benefits. Even small scandals are promoted in social media with more self-created stories, which breaks the self-confidence of individuals. Many reasons are involved in such false stories but mainly false stories are propagated because of money they obtain from advertisements. It is evident that, increase in fake stories creates chaos in society. Many cases have been reported where morphed videos or images are used to create disturbance in the society. So controlling such fake stories is a difficult job. This is where new technologies come handy. Machine learning and deep learning techniques have given many algorithmic concepts to address this problem. We used Passive-Aggressive Classifier to find fakeness in news stories. Naïve Bayes Classifier, Support Vector Machines (SVM), K-Nearest Neighbors, Decision tree, and random forest algorithms can also be used for the same purpose. However, Passive-aggressive algorithm provides better accuracy. So, it is used in our current study. We have two datasets, one from Kaggle Fake News available from kaggle.com and another dataset is from fact check website by name politifact.com. Few deep-learning algorithms like convolution neural networks and recurrent neural networks can also be used to solve false news detection. Artificial intelligence has proved success in the areas of voice and image recognition provides some efficient algorithms to solve this problem [1].

Associated Work

Lot of work has been done on automatic summarization of news articles. Such methods can condense a lengthy document to a summary using a custom web crawler and linguistic techniques like Triplet extraction, Semantic similarity calculation and OPTICS clustering [2]. With such data, we can get summary of the news article but we have to find whether the news is true not false. Discovering untrue stories that are flooded in social media is difficult task without proper dataset. The purpose of such analysis is to test a news article whether it is true or not. Machine learning and deep learning techniques help in identifying the fake stories. Initially, fake news detection was using language features of the reported data. Naïve Bayes classifier [3] was the simplest way suggested which uses language features in the reported data. Some deep learning models like Convolution Neural Network and Recurrent Neural Networks [4] are suggested to identify false news in twitter data. Even a combination of different machine learning algorithms like Support Vector Machines, K-Nearest Neighbors, Decision tree and Random forest [5] can be used to solve the problem. A cloud-based system was also proposed for identifying fake tweets using reverse image research, machine learning and source comparison [6]. Hierarchical attention networks are used to get radiant features in the text and find the validity of the text [7]. A new block chain system was proposed to control the extensive spread of fake news across network [8]. Combination of TF-IDF, word2vec and Support Vector Machine is used to identify fake news in big-data environment [9]. Irrespective of method, we have grave need of finding a stable, efficient and high accuracy model to define, detect and stop fake news and to preserve the authenticity of the news system. Therefore, in this paper, we have used Passive-Aggressive Classifier algorithm, which uses a defined method of classification and creates a labelled model. This labelled model is updated only when there is a wrong detection and in all other cases model parameters are untouched [10]. We have explored accuracy of the model and its ability to predict the trueness of news and the challenges faced to perform extensive testing.

Proposed Work

In recent times, machine learning and deep learning techniques has become antidote to every problem. Many problems that are difficult and time consuming are solved using machine learning techniques in an efficient way reducing both manual and computational works. Passive Aggressive Classifier is a supervised learning technique. It requires dataset to train and create a model. This data should undergo a lot of sanitization before being fed to machine learning algorithm. One of the methods of supervised machine learning is classification. In this paper, we have used classification, which is a method of predicting the class of data objects. Classification groups data into selected categories. Passive-Aggressive Classifier is a classification technique which gives solution to the problem of fake news. In this application, the following steps are used.

A.Web Scrapping of Data

A part of required data for training was obtained from fact checking website, politifact.com using, pandas and beautifulsoup libraries of python. In politifact website, news post is labelled with different labels such as “true”, “mostly true”, “half true”, “false”, “mostly false” and “pants on fire”. Among these labels we considered only data marked with “true” as the real news and everything else is considered fake. Hence, we formed dataset from web scrapping of approximately 600 pages. We obtained around 19,000 records. In Later stages, data is read from internet to update the model continuously at regular intervals.

B. Dataset

News classification is performed on two datasets. Kaggle Fake News dataset from kaggle.com is one of them. In addition, a dynamic dataset is created and used along with this dataset by performing web scrapping on politifact.com as mentioned in the previous step. The datasets are combined for analysis. Around 97 MB of data is considered for analysis which consists of approximately Forty thousand records.

C. Dataset preparation

The stop words (such as “is”, “the”, “a”, “an”) are removed from the dataset. The accuracy of the model depends on data cleaning, prioritization, and understanding of the data.

D. Splitting data to train and test set for model creation

In supervised learning always, data is split into arbitrary train and test data. In our current experiment, training is done on different splits such as 75% (train set) and 25% (test set), 70% (train set) and 30% (test set), 80% (train set) and 25% (test set).

E. TF-TDF

It is combination of count vectorizer followed by tf-idf transformer. Count vectorizer gives bag of words, which is then normalized to tf-idf representation by tf-idf transformer. Term Frequency (TF) represents the numbers of times a word or term is repeated in a document. Ratio of Number of times a term repeated to the total length of the document. Inverse document frequency (IDF) is the logarithmic ratio of total number of documents in the corpus to the documents containing the term in it.

$$\text{tf-idf}(t, d) = \text{tf}(t,d) * \text{idf}(t,d)$$

$$\text{idf}(t,d) = \log [(1 + n) / (1 + \text{df}(t))] + 1$$

Where, t represents term, d represents document and n represents total number of documents in the corpus.

F. Use Passive Aggressive Classifier for classification

Passive-Aggressive algorithm is an online algorithm mainly used for dataset, which is large and is continuous. This algorithm remains idle when there is positive result saying its prediction is right, which signifies passive part of it. However, it becomes aggressive once there is a negative result and start making changes. Unlike other algorithms, it does not change much, it just do a little improvement over its weight vector position, which should help in correcting the error. Unlike batch learning, where complete training dataset is available at once, in online learning the data is continuous in multiple steps. The machine-learning model should adapt to the data and should be updated in multiple steps. The classifier chosen is able to handle large data, which cannot be trained at once. Passive Aggressive algorithm takes tf-idf matrix generated in the previous step and uses it to train the model. Later test set is used to predict the accuracy of the classifier. Even though, with split of train and test data overfitting can be avoided, it still exists in case of some hyper parameter settings for estimator. So a k-fold cross validation is performed in order to avoid overfitting. Fig. 1 shows how the model is trained using train and test data and how best parameters are taken from cross validation to improve the model.

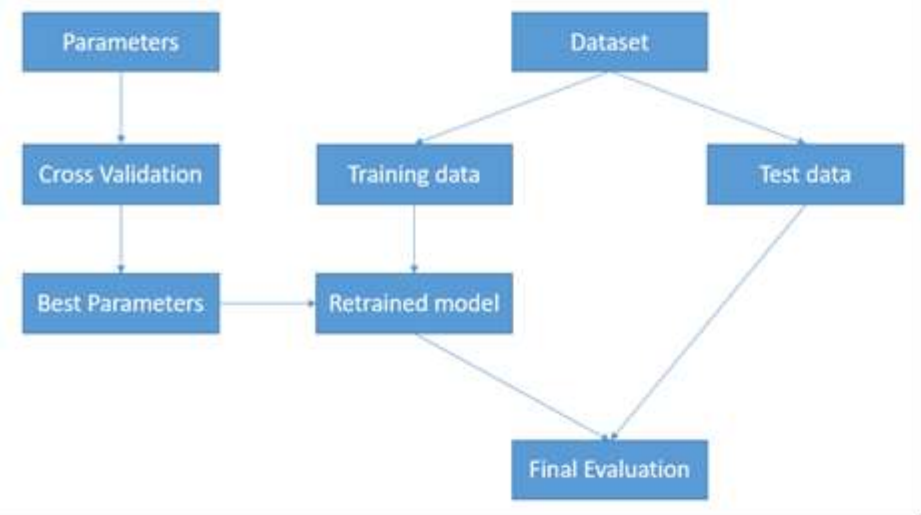
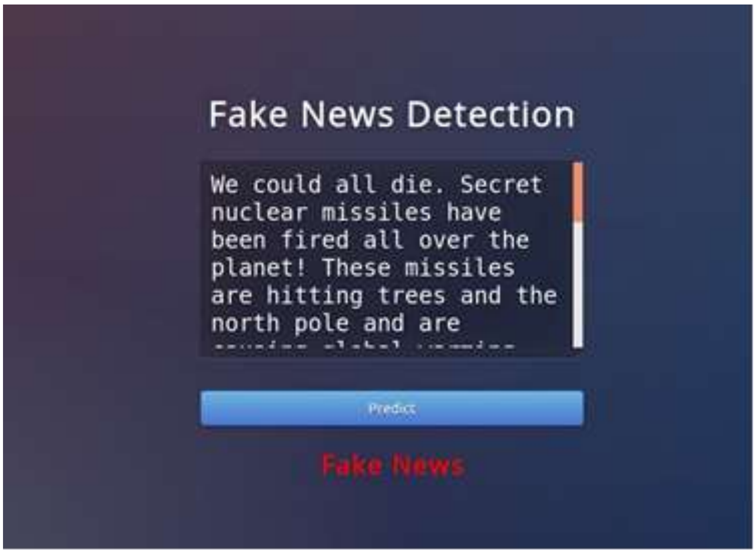


Fig 1: Process of training model using train and test data

Design of front-end interface

A flask based front-end interface is created to enter news and get the prediction of the model as shown in Fig. 2.



Fake News Detection

We could all die. Secret nuclear missiles have been fired all over the planet! These missiles are hitting trees and the north pole and are

Fake News

Fig 2: Front-end interface

Flow Diagram

Fig.3 shows the event flow diagram of the experiment. Web scrapped data from politifact.com and kaggle fake news dataset from kaggle.com are taken and data pre-processing is done by removing the stop-words. Then pre-processed data is split into train and test data. Both data is fed to TF-IDF vectorizer which converts it to TF-TDF feature matrices. Train data feature matrix is then fed to Passive aggressive classifier for training the model. Once model is trained, test data feature matrix is used to predict the accuracy. Later, the model gets updated whenever it receives data from the politifact.com. If user wants to know whether a news is valid, he can check it with the front end provided, by entering the news post in the text box and model predicts the validity of news.

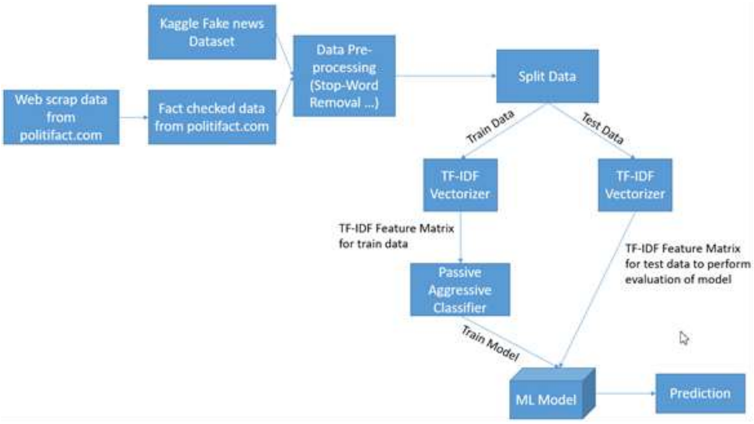

Kaggle Fake news Dataset
Web scrap data from politifact.com
Fact checked data from politifact.com
Data Pre-processing (Stop-Word Removal ...)
Split Data
Train Data
Test Data
TF-IDF Vectorizer
TF-IDF Vectorizer
TF-IDF Feature Matrix for train data
Passive Aggressive Classifier
TF-IDF Feature Matrix for test data to perform evaluation of model
Train Model
ML Model
Prediction

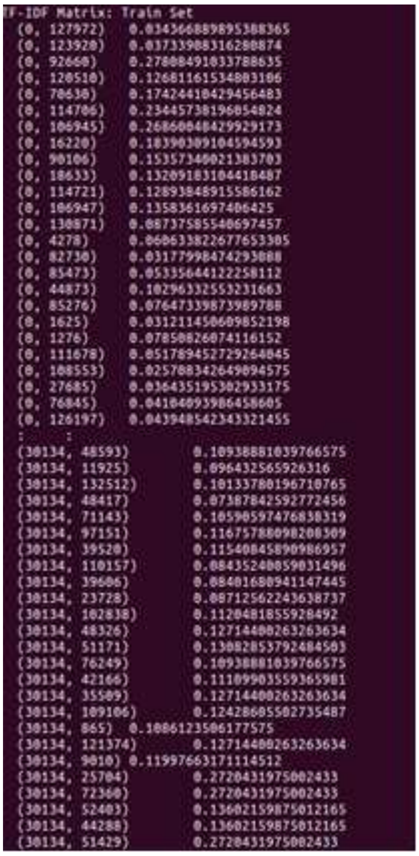
Fig. 3 Event flow diagram

3. Results and Discussion

Table 1: Results For Various Ratio Of Data Splits

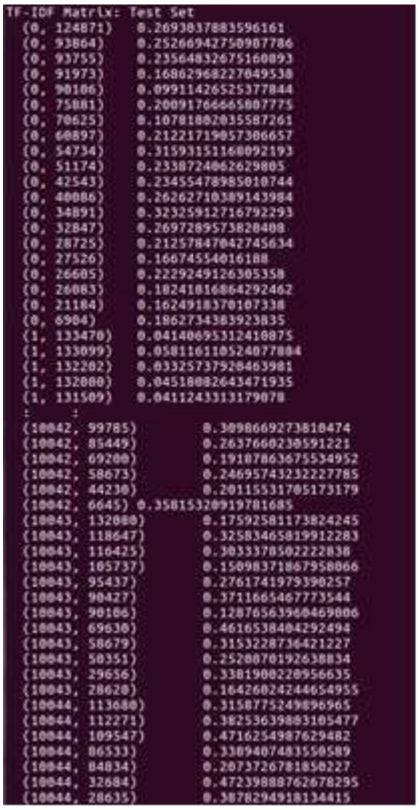
Dataset	Methodology	Train (%)	Test (%)	Accuracy (%)	K-Fold Cross Validation (%)
Kaggle	Passive Aggressive Classifier	70	30	96.35	95.947505 96.091158 95.773835
		75	25	95.67	96.149409 95.975768 95.975768
		80	20	95.94	95.990770 95.975768 95.845954
Politifact		70	30	83.59	84.878502 79.798762 75.228364
		75	25	79.65	83.501006 79.195046 74.856789
		80	20	81.79	84.398700 78.312693 77.550704

After the construction of model on Kaggle and Politifact datasets individually, the results have been documented in Table.1 for different splits of train and test data. For combined dataset, a model is constructed and evaluated on test data. Combined dataset has27389 fake news records and 12791 real news records. For a data split containing 75% of training data and 25% of testing data, TF-IDF feature matrix for train data is as shown in Fig. 4. This matrix is fed to Passive Aggressive Classifier for model creation. TF-IDF feature matrix for test data is shown in Fig. 5.



```
TF-IDF Matrix: Train Set
  (0, 127972)   0.03436688989538365
  (0, 123920)   0.03733908316280874
  (0, 92660)    0.27808491033788635
  (0, 120510)   0.12681161534803106
  (0, 70630)    0.17424410429456483
  (0, 114706)   0.23445738196054824
  (0, 106945)   0.26860048429929173
  (0, 16220)    0.18390309104594593
  (0, 90106)    0.15357340021383703
  (0, 18633)    0.13209183104418487
  (0, 114721)   0.12893848915586162
  (0, 106947)   0.1358361697406425
  (0, 130871)   0.08737585540697457
  (0, 4278)     0.060633822677653305
  (0, 82730)    0.03177998474293088
  (0, 85473)    0.05335644122258112
  (0, 44873)    0.10296332553231663
  (0, 85276)    0.07647339873989788
  (0, 1625)     0.031211450609852198
  (0, 1276)     0.07850826074116152
  (0, 111678)   0.051789452729264045
  (0, 108553)   0.025708342649094575
  (0, 27685)    0.036435195302933175
  (0, 76845)    0.04104093986458605
  (0, 126197)   0.043948542343321455
  :     :
  (30134, 48593)        0.10938881039766575
  (30134, 11925)        0.096432565926316
  (30134, 132512)       0.10133780196710765
  (30134, 48417)        0.07387842592772456
  (30134, 71143)        0.10590597476838319
  (30134, 97151)        0.11675788098208309
  (30134, 39520)        0.11540845890986957
  (30134, 110157)       0.08435240059031496
  (30134, 39606)        0.08401680941147445
  (30134, 23728)        0.08712562243638737
  (30134, 102838)       0.1120481855928492
  (30134, 48326)        0.12714400263263634
  (30134, 51171)        0.13082853792484503
  (30134, 76249)        0.10938881039766575
  (30134, 42166)        0.11109903559365981
  (30134, 35509)        0.12714400263263634
  (30134, 109106)       0.12428605502735487
  (30134, 865)  0.1086123506177575
  (30134, 121374)       0.12714400263263634
  (30134, 9010) 0.11997663171114512
  (30134, 25704)        0.2720431975002433
  (30134, 72360)        0.2720431975002433
  (30134, 52403)        0.13602159875012165
  (30134, 44288)        0.13602159875012165
  (30134, 51429)        0.2720431975002433
```

Fig 4: TF-TDF feature matrix of train data



```
TF-IDF Matrix: Test Set
  (0, 124871)   0.2693037883596161
  (0, 93864)    0.25266942750907786
  (0, 93755)    0.23564832675160893
  (0, 91973)    0.16862968227049538
  (0, 90106)    0.09911426525377844
  (0, 75881)    0.20091766665807775
  (0, 70625)    0.10781802035587261
  (0, 60897)    0.21221719057306657
  (0, 54734)    0.31593151168092193
  (0, 51174)    0.2338724062629805
  (0, 42543)    0.23455478985010744
  (0, 40086)    0.26262710389143984
  (0, 34891)    0.32325912716792293
  (0, 32847)    0.2697289573828408
  (0, 28725)    0.21257847042745634
  (0, 27526)    0.16674554016188
  (0, 26605)    0.2229249126305358
  (0, 26083)    0.18241016864292462
  (0, 21184)    0.1624918370107338
  (0, 6904)     0.1862734383923835
  (1, 133470)   0.04140695312410875
  (1, 133099)   0.058116110524077884
  (1, 132202)   0.03325737920463981
  (1, 132080)   0.045180082643471935
  (1, 131509)   0.0411243313179078
  :     :
  (10042, 99785)        0.3098669273810474
  (10042, 85449)        0.2637660230591221
  (10042, 69200)        0.19107863675534952
  (10042, 58673)        0.24695743232227785
  (10042, 44230)        0.20115531705173179
  (10042, 6645) 0.35815320919781685
  (10043, 132080)       0.17592581173824245
  (10043, 118647)       0.32583465019912283
  (10043, 116425)       0.3033378502222838
  (10043, 105737)       0.15098371867958066
  (10043, 95437)        0.2761741979390257
  (10043, 90427)        0.3711665467773544
  (10043, 90106)        0.12876563960469006
  (10043, 69630)        0.4616538404292494
  (10043, 58679)        0.3153228736421227
  (10043, 50351)        0.2520870192638834
  (10043, 29656)        0.3381900220956635
  (10043, 28620)        0.16426024244654955
  (10044, 113680)       0.3158775249896965
  (10044, 112271)       0.38253639883105477
  (10044, 109547)       0.4716254987629482
  (10044, 86533)        0.3309407483550589
  (10044, 84834)        0.2073726781850227
  (10044, 32684)        0.47239888762678295
  (10044, 28635)        0.3878294918134415
```

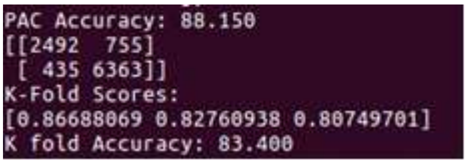
Fig 5: TF-TDF feature matrix of test data

Test data feature matrix is used to evaluate model and the accuracy is 88.66%. Confusion matrix for combined data is shown Table.2. A confusion matrix is a table, which visualizes the performance of supervised learning. Accuracy is 88.66%. Accuracy= (TruePostive + TrueNegative) / TotalSample = 88.66%.

Table 2: Confusion Matrix

Confusion Matrix	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positive = 2454	False Negative = 733
Predicted Negative (0)	False Negative = 406	True Negative = 6452

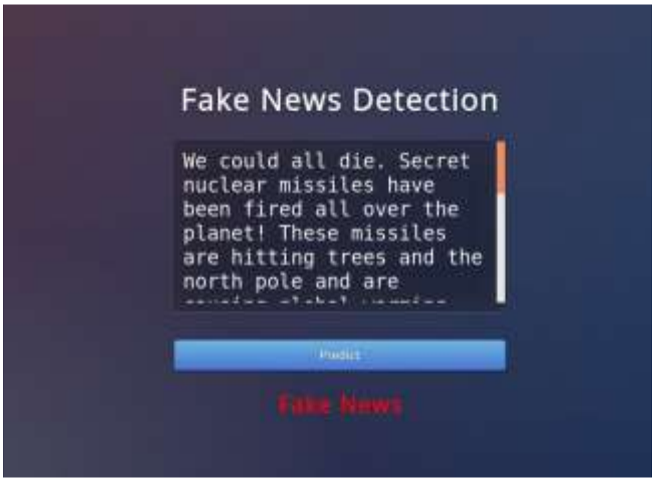
A K-fold cross validation is performed and accuracy is calculated in different folds of data and a mean K-Fold accuracy is calculated. Mean of K-Fold accuracies is 83.46% as shown in Fig. 6.



```
PAC Accuracy: 88.150
[[2492  755]
 [ 435 6363]]
K-Fold Scores:
[0.86688069 0.82760938 0.80749701]
K fold Accuracy: 83.400
```

Fig 6: PAC Accuracy and K-Fold Accuracy

Model is evaluated from the frontend created. A news statement is entered in the provided text box and the model's prediction is provided for the user as shown in [Fig: 7] & [Fig: 8].

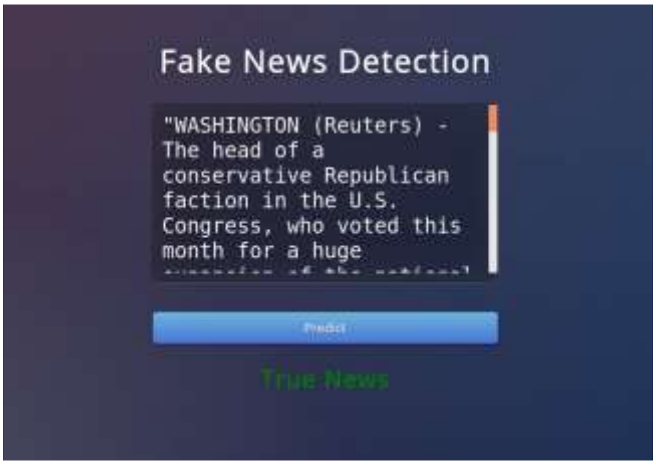


Fake News Detection

We could all die. Secret nuclear missiles have been fired all over the planet! These missiles are hitting trees and the north pole and are

Fake News

Fig 7: News predicted as fake on front-end



Fake News Detection

"WASHINGTON (Reuters) - The head of a conservative Republican faction in the U.S. Congress, who voted this month for a huge

True News

Fig 8: News predicted as true on front-end

4. Conclusion

Fake news creates negative influence on the society. It leverages weakness in the society of believing the things without inspection and make people suffer from fake and morphed information. It is high time to fight against fake news as such posts have increased in alarming rate. A need of using new technologies such as machine learning, deep learning and artificial intelligence to fight against fake news can yield some good result and protect the society. One such effort is done by using passive aggressive algorithm in this paper. As most of such fake posts are from online, Passive Aggressive Classifier algorithm, which is an online algorithm, plays an important role in avoiding fake news spread. These algorithms are powerful to identify/fight against the fake news propaganda.

References:

[1] Metz C (2016) “The bittersweet sweepstakes to build an AI that destroys fake news”, Dec 2016 (Online).

[2] Laxmi B.Rananavare, P.Venkata Subba Reddy, “Automatic News Article Summarization”, International Journal of Computer Sciences and Engineering, Vol.6, Issue.2, pp.230-237, 2018

[3] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), 2017, pp. 900-903, doi: 10.1109/UKRCON.2017.8100379.

[4] Ajao, Oluwaseun, Deepayan Bhowmik, and Shahrzad Zargari. "Fake news identification on twitter with hybrid cnn and rnn models." Proceedings of the 9th international conference on social media and society. 2018

[5] Lakshmanarao, A., Y. Swathi, and T. Srinivasa Ravi Kiran. "An efficient fake news detection system using machine learning." Int J Innov Technol Exploring Eng (IJITEE) 8.10 (2019).

[6] Krishnan, Saranya, and Min Chen. "Cloud-Based System for Fake Tweet Identification." 2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM). IEEE, 2019.

[7] Ahuja, Nishtha, and Shailender Kumar. "S-HAN: Hierarchical Attention Networks with Stacked Gated Recurrent Unit for Fake News Detection." 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO). IEEE, 2020.

[8] Saad, Muhammad, Ashar Ahmad, and Aziz Mohaisen. "Fighting fake news propagation with blockchains." 2019 IEEE Conference on Communications and Network Security (CNS). IEEE, 2019.

[9] Zhang, Shenhao, Yihui Wang, and Chengxiang Tan. "Research on text classification for identifying fake news." 2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC). IEEE, 2018.

[10] Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., & Singer, Y. (2006). Online passive aggressive algorithms.