Smart and Efficient Fake News Detection using Linguistic and Blog Based Dataset

Jayendra Kumar[1], Anumeha[2], Arvind R.Yadav[3]and M. Ramesh Naik[4]

Department of Electronics & Communication Engineering

NIT Jamshedpur, India[1, 4], Govt. Women's Polytecnic Jamshedpur[2], Parul University, Vadodara, India[3]

***E-mail:-** jkumar.ece@nitjsr.ac.in[1],rameshnaik466@gmail.com[2], amehanitjsr@gmail.com[3], arvind.yadav.me@gmail.com[4]*

DOI:- https://doi.org/10.47531/MANTECH/ECC.2021.39

Abstract

Fake News is false information about any existing original news content or intentionally fabricated for any specific purpose. Since the spread of news is more being used in an online manner, it's challenging to detect fake news automatically before it leads to any serious damage. Many researches have been done to differentiate between fake or real news content, using different dataset and algorithms. We introduce a comparative experiment on various classification algorithms and develop an efficient machine learning model to detect whether the news is fake or real. The experiment is done on two different formats of the dataset, which are mostly affected by fake news content, i.e., blog based (Face- book post) and linguistic-based news content. Thus, developing an efficient model with high accuracy and detecting the veracity of news in different format of news. We achieve an accuracy of 97.5% with the linguistic dataset and 75.5% with the Facebook post-based dataset.

Keywords: - *Machine learning, Fake news detection, Classifier algorithm, Facebook post, Linguistic news.*

INTRODUCTION

Fake news detection is a wide area topic to be researched and identify whether a news article is fake or real. Its area of research includes the various domain of news article and news article on the internet from various internet sources and social media [1-13]. The fast-growing internet and dependency on online news and article content have made it necessary to make a machine model, which can identify the veracity of any news content. According to the survey, it is found that the variety of news and its domain is wide to take all together to detect all type of news, and though it is impossible to detect a difference in fake or real news by human observation, so it is desired to develop a machine to automatically detect fake news from any type of news content. Some existing work has largely focused on detecting fake news based on their linguistic pattern, image quality, video quality etc. But the main source of news is always the linguistic content type of news, and a new form of news online is through social media posts.

In this paper, we introduce an automatic machine learning-based fake news detection technique, using Facebook post-based content and sports news dataset. These datasets are collected directly from an online resource. Thus, using these datasets, we build a fake news detector with an accuracy of 75.5% for Facebook post dataset and 97.5%.

RELEATED WORK

Fake news detection work is of many types, such as rumour, spam, hoaxes, etc. Many ideas have been proposed, like based on user linguistic content, image-based, video-based, based on the propagation of news around the internet and social media content.

Fake news is a growing area of research interest topic these days, and many researches have also been done in an attempt to detect the fake news automatically and efficiently. The major work focus on the detection of fake news is using a linguistic pattern of the news article content. Some paper also attempted to use image-based detection to make the model more efficient and cover a major area of the article to detect any fake news.

Zhiwei Jin, in [1] has proposed such image and content-based technique to detect the fake news in microblogs on the internet, as they supposed that many news blogs contain an article with an image attached which depicts major information of news article.

PrakharBiyani, in [2], has proposed a different way to detect news, i.e., by identifying the presence of Click baits in any news article. Click baits are some link designed, which invoke users to click on the link, which may lead to some non-profit content to the user. MykhailoGranik proposed a linguistic-based approach using a Naïve Bayes classifier, which successfully detects fake news in any news content article.

Cody Buntain, in [3], has proposed a way to detect fake news in Twitter news threads datasets collected from news sources. Kai Shu, in [4], has proposed a data mining view to detect any fake news by mining the information from any news article content and verifying thus whether the news article has truly happened or not.

The results so far obtained by different research work: such as 74% using linguistic news content by Naïve Bayes and Twitter post-based dataset with image features achieved 83.6% of accuracy.

Our work achieves an accuracy of 97.5% with stochastic gradient descent for the linguistic content dataset and 75.5% with a Facebook post-based dataset with random forest.

Many feature extraction technique has been already used in the different domain by several authors [14- 19].

DATA SET

The dataset used in this work is taken from online resources. One of the datasets isa Facebook fact check post dataset, and another one is the linguistic article dataset of the sports domain of news. These datasets are labelled.

A) Facebook fact check data set

The Facebook post-based dataset was collected from an online resource [5], the BuzzFeed News dataset. This dataset is used for the training of the model and testing by dividing the dataset for both separately. This dataset contains information about the Facebook post, where each post represents one form of news article posted on Facebook pages. The posts are collected from three large Facebook page categories, each from the left, right, and mainstream. The large contribution for the dataset is taken from ABC News Politics, CNN Politics, Politico. All these 9 Pages of news source has well-reputed credibility from the Facebook platform.

These BuzzFeed fact-check news dataset employees' fact-checked posts during the posting of seven days (weekdays) from 19th September 2016 to 27th September 2016. Then they rated this post as four labels, "mostly true," "mostly false", "mixture of true and false" and "no factual content". In the dataset, for each post, they also collected the number of "share count", "reaction count", and "comment count". The dataset also contains the type of post, i.e., link, photo, text, video. The news post is collected from 9 different account id and 16 different post ids.

BuzzFeed team collected a total of 2282 news article like this, in which 1145 article were from the mainstream page and 666 from the right-wing page and 471 collected from the left-wing page. Table 1 shows The Facebook post-based dataset.

Table 1: Class distribution for Facebook post-based dataset

aset	Class	Entries
Facebook fact check (BuzzFeed news)	Mostly True	1669
	No factual Content	264
	Mixture of true and False	245
	Mostly False	104

B) Linguistic data set

The Linguistic dataset was collected from an online resource. This dataset contains news in text content, the headline, and the body of news mainly. This dataset is collected from many news sources pages, mainly from BBC, NY Times, CNN, news pages [6].

From the news page, each news data is collected as a combination of headline and body of news. The news article mainly consists of news from the sports domain of news along with some close related news to sports. This model thus detects the veracity of sports news mainly. The news articles were then labelled as fake or as "0" or "1".

The dataset contains a total of 4048 news article. In which 1872 data is real news, and 2137 is fake news content. The feature extraction of the text-linguistic dataset is done by using the "Count Vectorizer" text feature extractor module of Scikit-learn. This extractor produces output in the form of a sparse matrix of the count of words.

We can also do "Tfidf transformer" or "Hashing Vectorizer", which are also an efficient way of extracting features from text data. Parameters like 'stop_words' used to reduce the number of words, which gives less information about the classification of news [3].

More parameters can be used, such as ngram_range, vocabulary, encoding, stemming for similar words in the article. Table 2 shows The Linguistic Dataset Real vs Fake.

Table 2: Sample fake or real news from linguistic dataset

Real	Fake
“there is no specific treatment for disease caused by novel coronavirus and the Chinese government is heavily promoting traditional medicines as treatment for COVID-19”	"Chinese are not taking any medicine or any vaccine for corona virus. Every household has a corona virus case. They have stopped going to hospital for cure. They instead kill the virus with heat”

IMPLEMENTATION

The dataset, which is available in comma separated file format. Two datasets are to be loaded, “Facebook fact check dataset” and “Linguistic dataset”.

In the Facebook dataset, among 12 fields (account_id, post_id, Category, Page, Post URL, Date Published, Post Type, Rating, Debate, share_count, reaction_count, comment_count), we have used 5 field (account_id, post_id, share_count, reac- tion_count, comment_count) as feature for our models.

In the Linguistic dataset, we have used the news Body field for feature extraction and applying feature value to the model.

Thus, for the Facebook dataset, we have 4 labels, “Mostly True”, “mixture of true or false”, “No factual content”, “Mostly false”. And with the linguistic dataset, we have 2 labels, “1” indicating Real news and “2” indicating Fake news.

Data pre-processing and feature extraction is done to remove or replace null values from the dataset, make data fit for the model to be trained. For the text data, count vectorizer is used as the feature extractor to convert the text to numerical data.

The dataset is shuffled and split into training and test dataset, 80% for the training of model and 20% for the testing of model.

Training data is used to train the model with the features to value the data have, and the testing data is used for the estimation of how well the classifier algorithm performs on the new data.

Then we apply the Machine learning algorithms to our training and testing dataset.

For the model with the linguistic dataset, it uses the word counts in a fake and real news article; thus, if the number of word counts for the test data is more in fake labelled data, then it classifies the test data as fake; similarly, it classifies for fake.

Tuning of parameters for all algorithms used is done separately, and the classifiers are tuned with various parameters value to analyse the best parameter for the model.

Finally, the whole algorithm is run, and the average accuracy value from all classifier is taken into account to compare the best classifier for the respective dataset.

The same is shown by the flow graph in Fig 1 below.

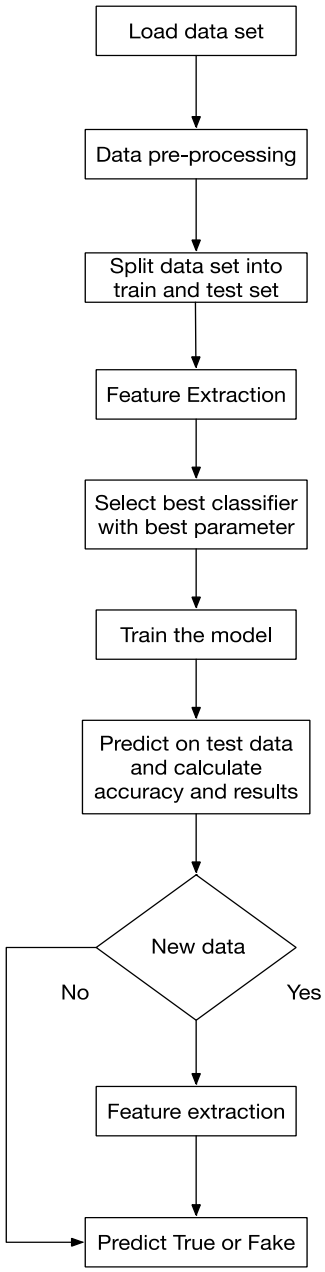

Fig 1: Proposed Algorithm flow for fake news detection

RECEIVED RESULTS

The result for the classifier is shown as accuracy chart, confusion matrix [14-17], classification report, accuracy plots, precision and recall.

Let's assume our result as positive, i.e. if the result gives the news is fake, then it indicates to be positive So,

The number of true positive (TP) indicates, number of actual fake news predicted as fake news.

The number of falsepositive (FP) indicates, number of actual real news predicted as real news.

The number of true negative (TN) indicates, number of actual fake news predicted as real news.

The number of false negative (FN) indicates, number of actual real news predicted as fake news.

The "precision" and "recall" of the classifier is calculated as follows: Precision =*tp/(tp + fp)*

Recall=*tp/(tp + fn)*

Where, tp = number of true positive example fp = number of false positive example fn = number of false negative example.

Figure 2 shows the confusion matrix of Fake Real news.

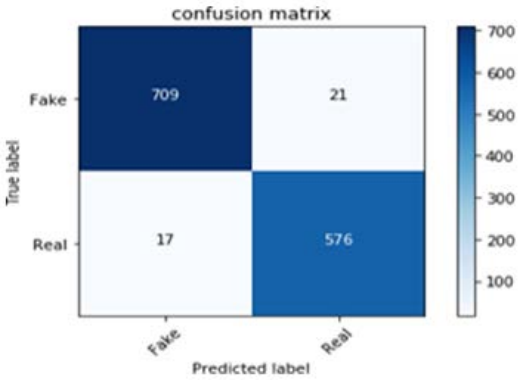

Fig. 2: Confusion Matrix

The accuracy of different classifier for linguist data and facebook post data has been shown in Table 3 and in the form of Bar-chat in Fig 3.

Table 3: Accuracy chart for all the classifier

S. No.	Algorithm	Linguistic data (%)	Facebook post data (%)
1	Random Forest	89.387	75.759
2	Support Vector Machine	72.144	72.428
3	Logistic Regression	97.356	72.866
4	K Nearest Neighbor	74.899	72.165
5	Decision Tree	94.890	66.092
6	Stochastic Gradient Decent	97.506	64.697
7	Naïve Bayes	96.9	27.789
8	Ensemble method	93.765	74.398

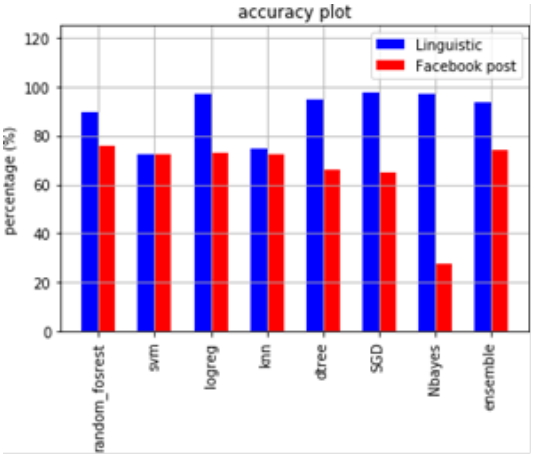

Fig 3: Accuracy plot of classifiers

CONCLUSION

Next, in the task of fake news detection, we observed the Facebook dataset and linguistic dataset. For different type of features, we developed multiple classifier algorithm and made the comparison for the same. We have used random forest, support vector machine, Logistic Regression, K Nearest neighbour, Decision tree, Stochastic gradient descent, Naïve Bayes and ensemble of all the above algorithms. Comparing all the algorithms accuracy and performance, we came to the conclusion that Stochastic gradient descent was better with text-based dataset giving average test accuracy of 97.5%, and with the Facebook post dataset, the random forest was comparatively good with an average test accuracy of 75.759%. The model is also developed for user interface with development of GUI, for checking the veracity of news article at real-time by any user.

The future work involves developing a dataset with multi-domain news and making a classifier for the same. Also, the work can be extended with an incremental way of detecting the fake or real from live streaming news articles from news pages online in real-time.

REFERENCES

1. Jin, Z., Cao, J., Zhang, Y., Zhou, J., &Tian, Q. (2016). Novel visual and statistical image features for microblogs news verification. IEEE transactions on multimedia, 19(3), (pp. 598-608).
2. Biyani, P., Tsioutsiouliklis, K., &Blackmer, J. (2016, February). "8 Amazing Secrets for Getting More Clicks": Detecting Clickbaits in News Streams Using Article Informality. In Thirtieth AAAI Conference on Artificial Intelligence.

3. Buntain, C., &Golbeck, J. (2017, November). Automatically identifying fake news in popular Twitter threads. In 2017 IEEE International Conference on Smart Cloud (SmartCloud) (pp. 208-215).
4. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), (pp. 22-36).
5. Wu, L., Li, J., Hu, X., & Liu, H. (2017, June). Gleaning wisdom from the past: Early detection of emerging rumors in social media. In Proceedings of the 2017 SIAM international conference on data mining (pp. 99-107).
6. Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648. \
7. Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., & Krause, A. (2018, April). Fake news detection in social networks via crowd signals. In Companion Proceedings of the The Web Conference 2018 (pp. 517-524).
8. Della Vedova, M. L., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M., & de Alfaro, L. (2018, May). Automatic online fake news detection combining content and social signals. In 2018 22nd Conference of Open Innovations Association (FRUCT) (pp. 272-279).
9. Ruchansky, N., Seo, S., & Liu, Y. (2017, November). Csi: A hybrid deep model for fake news detection. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 797-806).
10. Gilda, S. (2017, December). Evaluating machine learning algorithms for fake news detection. In 2017 IEEE 15th Student Conference on Research and Development (SCOReD) (pp. 110-115).
11. Granik, M., &Mesyura, V. (2017, May). Fake news detection using naive Bayes classifier. In 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON) (pp. 900-903).
12. Castillo, C., El-Haddad, M., Pfeffer, J., &Stempeck, M. (2014, February). Characterizing the life cycle of online news stories using social media reactions. In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (pp. 211-223).
13. Pérez-Rosas, V., Kleinberg, B., Lefevre, A., &Mihalcea, R. (2017). Automatic detection of fake news. arXiv preprint arXiv:1708.07104.
14. Kumar, J., Anand, R.S. &Srivastava, S.P. (2014 Jan). Multi-class welding flaws classification using texture feature for radiographic images. In procedding of International Conference on Advances in Electrical Engineering (ICAEE). IEEE. (pp.1-4)
15. Kumar, J, Anand, R.S., &Srivastava, S.P., (2014, Feb). Flaws classification using ANN for radiographic weld images. In procedding of International Conference on Signal Processing and Integrated Networks (SPIN). IEEE (pp. 145-150).
16. Kumar, J., Srivastava, S.P., Anand, R.S., Arvind, P., Bhardwaj, S. & Thakur, A., (2018, December). GLCM and ANN based Approach for Classification of Radiographics Weld Images. In procedding of 13th International Conference on Industrial and Information Systems (ICIIS) IEEE. (pp. 168-172).
17. Kumar, J., Arvind, P., Singh, P., Sarada, Y., Kumar, N. &Bhardwaj, S., (2019, December). LBP riu2 Features for Classification of Radiographic Weld Images. In Procedding of International Conference on Innovative Trends and Advances in Engineering and Technology (ICITAET) IEEE. (pp. 160-165).
18. Yadav, A.R., Anand, R.S., Dewal, M.L., Gupta, S.& Kumar, J., (2018). Comparison of feature extraction techniques for classification of hardwood species. International Journal of Computational Systems Engineering, 4(2-3), pp.106-119.
19. Yadav, A.R., Kumar, J., Anand, R.S., Dewal, M.L. & Gupta, S., (2018, March). Binary Gabor pattern feature extraction technique for hardwood species classification. In proceeding of 4th International Conference on Recent Advances in Information Technology (RAIT) IEEE. (pp. 1-6).