Fake News Detection using Machine Learning and Natural Language Processing

Kushal Agarwalla, Shubham Nandan, Varun Anil Nair, D. Deva Hema

***Abstract*: *The web and internet-based life have led the entrance to news data, a lot less demanding and agreeable. Mass-media affects the life of the general public and as it frequently occurs. There are few individuals that exploit these privileges. This prompts the creation of the news articles that are not totally evident or indeed, even totally false. People intentionally spread these counterfeit articles with the help of web-based social networking sites. The fundamental objective of fake news sites is to influence the popular belief on specific issues. The main goal of fake news websites is to affect public opinion on certain matters. Our aim is to find a reliable and accurate model that classifies a given news article as either fake or true*.**

Index Terms*: *Classification algorithm , Fake news detection, Machine learning, Natural language processing

I. INTRODUCTION

Modern life has become quite suitable and the people of the world have to thank the vast contribution of the internet technology for transmission and information sharing. There is no doubt that internet has made our lives easier and access to surplus information viable.

This is an evolution in human history, but at the same time it unfocusses the line between true media and maliciously forged media. Today anyone can publish content – credible or not – that can be consumed by the world wide web. Sadly, fake news accumulates a great deal of attention over the internet, especially on social media. People get deceived and don't think twice before circulating such mis-informative pieces to the far end of the world. This kind of news vanishes but not without doing the harm it intended to cause.

The given social media sites that play a major role in supplying counterfeit news include Facebook, Twitter, Whatsapp etc.

Many scientists believe that counterfeited news issue may be addressed by means of machine learning and artificial intelligence. This is because recently artificial intelligence algorithms have begun to improve work on lots of classification problems (image recognition, voice detection and so on) because hardware is cheaper and bigger datasets are available.

Various models are used to provide an accuracy range of 60-75%. Which comprises of Naïve Bayes classifier, Linguistic features based, Bounded decision tree model, SVM etc. The parameters that are taken in consideration do not yield high accuracy. The motive of this paper is to increase the accuracy of detecting fake news more than the present results that are available. An algorithm have been explored that can distinguish the difference between the fake and true news with an 83 percent accuracy. By fabricating this new model which will judge the counterfeit news articles on the basis of certain criteria which are as follows: spelling mistake, jumbled sentences, punctuation errors etc.

II. RELATED WORK

[1] identifies different media sources and analyses whether the given news article is credible or not. The paper provides with an insight on characterization of news article combined with different content types available. The paper uses models such as linguistic features based models and predictive modelling, which aren't up to par with the rest of the models present.

[2] predicted fake news through naïve Bayes classifier. This approach was implemented as a software system and tested against various data set of Facebook etc. which provided an accuracy of 74%. The paper did not consider punctuation errors, leading to a low accuracy.

[3] Evaluated different machine learning algorithms and analyzed the prediction percentage. The accuracy of different predictive models which included bounded decision trees, gradient boosting, and support vector machine were tabulated. The models are evaluated on the basis of probability threshold which aren't most reliable.

[4] discuss about fake news detection and ways to apply them on various Social media sites using naïve Bayes classifier. The data sources for news article are Facebook, twitter etc. The accuracy obtained is quite low as these site's information are not 100% credible.

[5] discusses about counteracting misinformation and rumour detection in real time. It uses novelty-based feature and attains its data source from Kaggle. The accuracy rate of the model is 74.5%. Clickbait and unreliable sources are not considered which led to lower accuracy.

[6] it is used to differentiate between spammers and non-spammers in Twitter. The various models used include naïve Bayes classifier, clustering and decision tree. Accuracy rate to detect spammers are at 70% and non-spammers are at 71.2%. The models that were used attained a low average accuracy to segregate spammer and non-spammer.

Revised Manuscript Received on March 25, 2019.

Kushal Agarwalla, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.

Shubham Nandan, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.

Varun Anil Nair, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.

D. Deva Hema, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.

[7] to detect fake news through various methods. The accuracy rate is limited to 76% as linguistic model was encouraged to be used. Higher accuracy can be obtained if predictive model were to be used.

[8] Detecting fake news through various machine learning models. The given machine learning models implemented are naïve Bayes classifier and support vector machine. No specific accuracy was recorded as only the models were discussed.

[9] to detect whether the given Tweets are credible or not. The machine learning model implemented are naïve Bayes classifier, decision trees, Support vector machines and neural networks. With both tweet and user features, the best F1 score is 0.94. Higher accuracy could have been attained by considering non-credible news into account.

[10] Method for automating fake news detection on Twitter by learning to predict accuracy assessments in two credibility-focused Twitter datasets. Accuracy rate of the given models are at 70.28%. The main limitation lies in the structural difference CREDBANK and PHEME, which could affect model transfer.

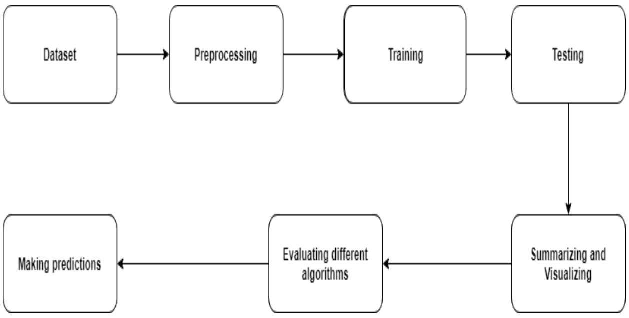

Fig 1. Pipeline Representation

III. Dataset and Feature Generation

A. Dataset Description

One of the most difficult issues to unravel in machine learning has nothing to do with complex calculations : it's the issue of getting the correct datasets in the correct organization. Getting the correct information implies assembling or distinguishing the information that relates with the results which needs to be foreseen; for example information that contains a flag about occasions which needs to be taken care about. The datasets should be lined up with the issue which is being attempted to explain. In the event that the correct information is not present, at that point the endeavours to assemble an AI arrangement must come back to the dataset gathering stage. Deep learning, and machine adapting all the more for the most part, needs a decent preparing set to work legitimately. Gathering and developing the training set – a sizeable assemblage of known information – requires some investment and area explicit learning of where and how to accumulate applicable data. The training set goes about as the benchmark against which machine learning nets are prepared. That is the thing which needs to be figured out how to remake before they're released on information they haven't seen previously. At this stage, educated people need to locate the correct crude information and change it into a numerical portrayal that the machine learning calculation can get it. Test sets that require much time or ability can fill in as an exclusive edge in the realm of information science and critical thinking.

Selecting the right dataset for Machine learning is very important to make the AI model functional with right approach. Though selecting the right quality and amount of data is challenging task but there are few rules needs to be followed for machine learning on big data.

In this project, the dataset is being taken from kaggle.com. The size of the dataset is 4008*4. It means that there are 4008 rows along with 4 columns. The name of the columns are “URLs”, “Headline”, “Body” and “Label” . It is being seen from the dataset that there are 2136 fake news articles and 1872 real news articles. Now it is to be seen the accuracy that the algorithm can provide.

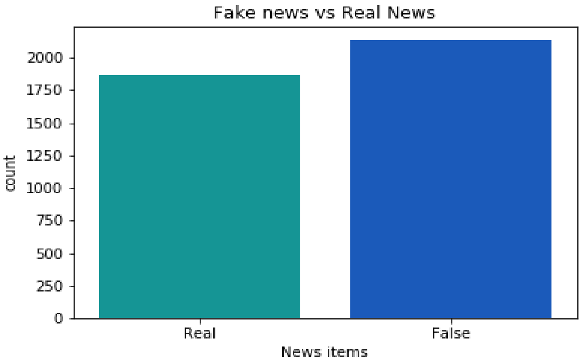

Fig 2. Fake and Real news article count

B. Preprocessing

Pre-preparing alludes to the changes connected to the information before nourishing it to the calculation. Data preprocessing is a method that is utilized to change over the crude information into a perfect informational index. At the end of the day, at whatever point the information is assembled from various sources it is gathered in a crude organization which isn't doable for the examination. Preprocessing is essential for accomplishing better outcomes from the applied model in Machine Learning project the configuration of the information must be in a legitimate way. Another perspective is that the dataset ought to be arranged so that more than one Machine Learning and Deep Learning calculations are executed in one informational index, and best out of them is picked. Before representing the data using various evaluating models, the data needs to be subjected to certain refinements. This will help us reduce the size of the actual data by removing the irrelevant information that exists in the data.

The data obtained was in csv format, and needed a lot of manual, syntactic preprocessing. A total of 4008 samples were distributed to train: test set in ratio 7:3. Each sample corresponds to a news article headline and body. NLTK in python was used to tokenize the body and headline. Removing the stop-words (referencing the NLTK stop-word list), helped in lemmatizing the rest of the data. To obtain the labeled sentence list for a particular course, the following processing steps were applied:

1. Tokenize the body and headline with the Punkt statement tokenizer from the NLTK library. This tokenizer runs an unsupervised machine learning algorithm pre-trained
on a

general English corpus, and can distinguish between sentence punctuation marks, and position of words in a statement.
2. Tag each sample with the tokens obtained from entire headline set, and body set.
A word cloud has been created for the headline and body present in our dataset. Word cloud is a novelty visual portrayal of content information, normally used to delineate catchphrase metadata (labels) on sites, or to imagine free structure content. Labels are typically single words, and the significance of each tag appears with text dimension or color. This design is valuable for rapidly seeing the most conspicuous terms and for finding a term in order to decide its relative unmistakable quality. At the point when utilized as site route helps, the terms are hyperlinked to things related with the tag.

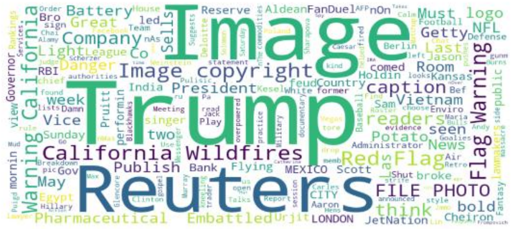

Fig 3. Word Count

C. Train and Test Splitting

To make a valuable training set, the issue needs to be comprehended for which it is being settled for. For example what will the machine learning calculation do and what sort of yield is anticipated. Machine learning regularly works with two informational collections: training and test. Each of the two ought to arbitrarily test a bigger assortment of information. The principal set which is being used is the training set, the biggest of the two. Running a training set through a machine learning system shows the net how to weigh diverse highlights, changing them coefficients as per their probability of limiting blunders in the outcomes. Those coefficients, otherwise called parameters, will be contained in tensors and together they are known as the model, since it encodes a model of the information on which it is being trained. They are the most vital takeaways which is being acquired from preparing a machine learning system. The second set is the test set. It works as a seal of endorsement, and is not utilized until the end. After it is being prepared and information is set, the neural net can be tested against this last arbitrary examination. The outcomes it produces ought to approve that the net precisely perceives pictures, or remembers them at any rate [x] level of them. On the off chance, that precise forecasts are not met, return to the training set and take a look at the mistakes made. Taking the right dataset would not create any kind of problem and the system will function smoothly.

IV. EVALUATION MODELS

A. Logistic Regression

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression which outputs continuous number values, logistic regression changes its yield utilizing the calculated sigmoid capacity to restore a likelihood esteem which would then be able to be mapped to at least two discrete classes. The LR model uses gradient descent to converge onto the optimal set of weights (θ) for the training set. For our model, the hypothesis used is the sigmoid function:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + exp(-\theta^T x)}$$

B. Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification and regression purposes. SVMs are mostly used in classification problems.

SVMs are founded on the idea of finding a hyperplane that best divides a dataset into two classes. Support vectors are the data points nearest to the hyperplane, the points of a data set that, if deleted, would alter the position of the dividing hyperplane. Because of this, they can be considered the critical elements of a data set. The distance between the hyperplane and the nearest data point from either set is known as the margin. The aim is to choose a hyperplane with the greatest possible margin between the hyperplane and any point within the training set, giving a higher chance of new data being classified correctly.

The expression for this kernel is given by the following expression:

$$G(x;\sigma) = \frac{1}{\sqrt[2]{2\pi\sigma}} exp(\frac{x^2}{2\sigma^2})$$

Note that this expression is provided for the 1-D case. In retrospect, the selection of this high-order kernel seems rather naive, since it may have caused the SVM model to over fit the training set.

C. Naïve Bayes Classification with Lidstone smoothing

In machine learning, Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with powerful (naive) independent assumptions between the features. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers. The formula for naïve bayes classifier is:

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)} = \frac{P(A) \cdot P(B|A)}{P(B)}$$

The Lidstone smoothing is a technique to smooth categorical data. A pseudo-count will be implemented in every probability estimate. This ensures that no probability will be zero. It is a way of regularizing Naïve Bayes. In general case, it is often called
as Lidstone smoothing.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Given an observation x = (x_1, …, x_d) from a multinomial distribution with N trials and parameter vector θ = (θ_1, …, θ_d), a "smoothed" version of the data gives the estimator:

$$\hat{\theta}_i = \frac{x_i + \alpha}{N + \alpha d} \qquad (i = 1, \ldots, d),$$

where the pseudo-count α > 0 is the smoothing parameter (α = 0 corresponds to no smoothing). Additive smoothing is a type of shrinkage estimator, as the resulting estimate will be between the empirical estimate xi / N, and the uniform probability 1/d. Most of the times, α is taken as 1 but a smaller value can also be chosen depending on the requirements.

The frequency-based probability might introduce zeros when multiplying the probabilities, leading to a failure in preserving the information contributed by the non-zero probabilities. Therefore, a smoothing approach, for example, the Lidstone smoothing, must be adopted to counter this problem.

After deciding on these problems, the Naïve Bayes classifier will be mostly used to obtain reasonable results. A smoothing approach will increase the accuracy of the problem which is being attempted. It is also being seen that Naïve Bayes classifier is both simple and powerful for Natural Language Processing (NLP) tasks such as text classification problems.

V. Result and Conclusion

Using the above-mentioned algorithms, i.e. Naïve Bayes classifier, Support Vector Machine and Logistic Regression, the following accuracy has been attained:

Feature Set	Naïve Bayes with lidstone smoothing	Support Vector Machine	Logistic Regression
Body+ Headline	0.8316	0.8165	0.6588
Body	0.8253	0.8165	0.6588
Headline	0.6805	0.6624	0.6657

The maximum accuracy of 83 percent on the given training set was attained by using Naïve Bayes classifier with lidstone smoothing. Whereas in the previous models which consisted of only Naïve Bayes (without lidstone smoothing) attained an accuracy of 74 percent.

The first algorithm used for classification was Naive Bayes (with Lidstone smoothing), where no hyper-parameter was required. This helped to set a reference point for further analysis. It was followed by SVM model where we selected the normalizing parameter (T) as 12. The model was trained starting from a smaller value of T = 4, because the larger the T the larger number of features influencing the output. However, the model did not converge for any T smaller than 12. Another hyper parameter used in SVM was Lagrange multiplier (λ). A λ value of 1/64 was used which gave the best result. Any value smaller than this was not converging.

The third model was Logistic Regression, where the only parameter used was learning rate (α). The learning rate between 5 to 12 was giving same convergence point, hence value of 10 was used. However, this model resulted in exceptionally low accuracy.

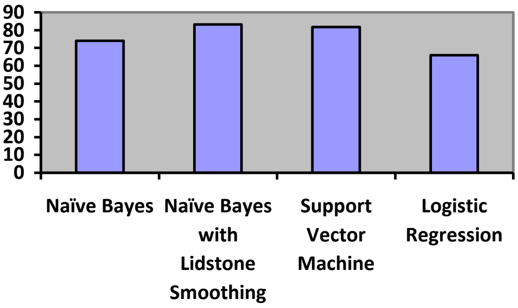

Fig 4. Comparison of accuracy attained by different evaluation models

VI. Future Work

A lot of our results circle back to the need for acquiring more accuracy. Generally speaking, simple algorithms perform better on less (less variant) data. Since we had a huge set of data, SVM, Naive Bayes and Logistic Regression underperformed. Given enough time to acquire more fake news data, and gain experience in python, we will try to better process the data using n-grams, and revisit our deep-learning algorithm. We tried using our own codes for the project, and the algorithms were relatively slow. To tweak all knobs of various algorithms, we shall use available robust packages in the future.

REFERENCES

1. Shivam B. Parikh and Pradeep K. Atrey, “Media-Rich Fake News Detection: A Survey”, IEEE Conference on Multimedia Information Processing and Retrieval, 2018.
2. Mykhailo Granik and Volodymyr Mesyura, “Fake News Detection Using Naive Bayes Classifier”, IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), 2017.
3. Shlok Gilda, “Evaluating Machine Learning Algorithms for Fake News Detection”, IEEE 15th Student Conference on Research and Development (SCOReD),2017.
4. Akshay Jain and Amey Kasbe, “Fake News Detection”, IEEE International Students' Conference on Electrical, Electronics and Computer Sciences, 2018.
5. QIN Yumeng, Dominik Wurzer and TANG Cunchen, “Predicting Future Rumours”, Chinese Journal of Electronics, 2018.
6. Arushi Gupta and Rishabh Kaushal, “Improving Spam Detection in Online Social Networks”, International Conference on Cognitive Computing and Information Processing(CCIP), 2015.
7. Veronica Perez-Rosas, Bennett Kleinberg, Alexandra Lefevre and Rada Mihalcea1, ”Automatic Detection of Fake News”, 2018.
8. Supanya Aphiwongsophon and Prabhas Chongstitvatana, ” Detecting Fake News with Machine Learning Method”, CP Journal, 2018.
9. Stefan Helmstetter and Heiko Paulheim,”Weakly Supervised Learning for Fake News Detection on Twitter”, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2018.
10. Cody Buntain and Jennifer Golbeck, “Automatically Identifying Fake News in Popular Twitter Threads”, IEEE International Conference on Smart Cloud (SmartCloud), 2017.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication