e-ISSN: 2582-5208

International Research Journal of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

FAKE NEWS DETECTION USING NAÏVE BAYES ALGORITHM

Manisha Moorpani[*1], Muskan Patel[*2], Bhagyashri Bharule[*3], Maher Quazi[*4]

[*1,2,3,4]Department Of Computer Engineering ,SSBT College Of Engineering And Technology,
Jalgaon, 425001, India.

ABSTRACT

In our era wherever the web is omnipresent, everybody depends on varied on-line resources for news. together with the rise within the use of social media platforms like Facebook, Twitter, etc. news unfolds chop-chop among ample users at intervals an awfully short span of your time. The unfold of faux news has sweeping consequences just like the creation of biased opinions to swaying election outcomes for the advantage of sure candidates. Moreover, spammers use appealing news headlines to come up with revenue mistreatment advertisements via click-baits. So, we tend to develop an NLP fake news detector by using python libraries like skit learn, matplotlib, pandas. We tend to area unit sleuthing faux news on US elections and we'll predict the news is faux or true by machine learning technique. The model that we tend to use is Multinominal Naive Bayes Classifier that offer 95% accuracy. And when predicting if the news is faux or true, we tend to use flask app to indicate results higher user interface expertise.

Keywords: Fake News, Prediction, Python, Classifier, Flask.

I. INTRODUCTION

We can get on-line news from completely different sources like social media websites, computer programmed, homepage of reports agency websites or the fact checking websites. On the net, there are a unit many publicly out datasets for pretend news classification like Buzzfeed News, BS Detector, Kaggle datasets, etc. As associate increasing quantity of our lives is spent interacting on-line through social media platforms, a lot of individuals tend to hunt out and consume news from social media rather than ancient news organizations. It's usually a lot of timely and fewer overpriced to consume news on social media compared with ancient journalism, like newspapers or tv, and it's easier to more share, discuss, and discuss the news with friends or different readers on social media. Our review analyses the way to discover the pretend news on social media to beat this downside. There are techniques to validate the style of the users to classify the news content, however these strategies even have their outliers and error rates. So, we've projected a system which will discover pretend news by machine learning exploitation python libraries.

II. METHODOLOGY

Dataset

The dataset that we had used is explained below:

The dataset is collected from Kaggle .It includes text data. The labels for news truthfulness are fine-grained multiple classes: pants-fire, false, barely true, half-true, mostly true, and true. The data source used for this project is 2016 US Elections dataset which contains file with .csv format for test and train. We divide the data into 80% real and 20% fake and named it as "fake_or_real_news.csv". It contains 4 columns viz ,

1. id: unique id for the news articles

2. title: the title of a news article

3. text: the text of the article; incomplete in some cases

4. Label: a label that marks the article reliable or unreliable

Cleaning and Pre-processing

Cleaning up the text information is important to spotlight attributes that we're about to need our machine learning system to select abreast of. improvement (or pre-processing) the info generally consists of many steps:

1. Remove Punctuation: Punctuation will offer grammatical context to a sentence that supports our understanding. except for our vectorizer that counts the amount of words and not the context, it doesn't add worth, thus we tend to take away all special characters. eg: How are you?->How are you

2. Tokenization: Tokenizing separates text into units like sentences or words. It provides structure to antecedently unstructured text.eg: Plata o Plomo-> 'Plata','o','Plomo'. After tokenization, we must make sure all tokens considered contribute to the label prediction.

3. Remove Stop words: Stop words are common words that will likely appear in any text. Stop words including words as "the", "as" and "and" appear a lot in a text, but do not give a relevant explanation. For this reason, they are removed.

4. Lemmatization: Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma .e.g., when took or taken are read in the text, they are lemmatized

NLP Techniques

We area unit exploitation Pos-tagging and TF-IDF for word frequency scores that attempt to highlight words that area unit a lot of fascinating, e.g., frequent in a very document however not across documents. Pos-tagging could be a method to price the words in text format for a selected a part of a speech supported its definition and context. it's chargeable for text reading in a very language and distribution some specific token (Parts of Speech) to every word. The TfIdf-Vectorizer can tokenize documents, learn the vocabulary and inverse document frequency weightings, and permit you to cypher new documents. whereas coaching the dataset this is often important step. Formula(Algorithm):

TF-IDF = TF(t,d) * IDF(t) '

TF' stands for term frequency and 'IDF' stands for Inverse document frequency ,in this 't' is a term frequency that is number of times t appears in a doc,'d'.

IDF is calculated by : log[1+n/1+df(d,t)]+1

III. MODELING AND ANALYSIS

Naïve Bayes Classifier for Fake News Detection: This classification technique is based on Bayes theorem, which assumes that the presence of a particular feature in a class is independent of the presence of any other feature. It provides way for calculating the posterior probability.

$$P(A|B) = P(B|A) \cdot P(A) / P(B)$$

Finding the probability of event, A when event B is TRUE

P (A) = PRIOR PROBABILITY

P (A|B) = POSTERIOR PROBABILITY FINDING PROBABILITY:

P (A|B1) = P (A1||B1). P (A2||B1). P (A3||B1)

P (A|B2) =P (A1||B2). P (A||B2).

P (A3||B2)

If the probability is 0 P (Word) = Word count +1/ (total number of words+ No. of unique words) Therefore, by using this formula one can find the accuracy of the news.

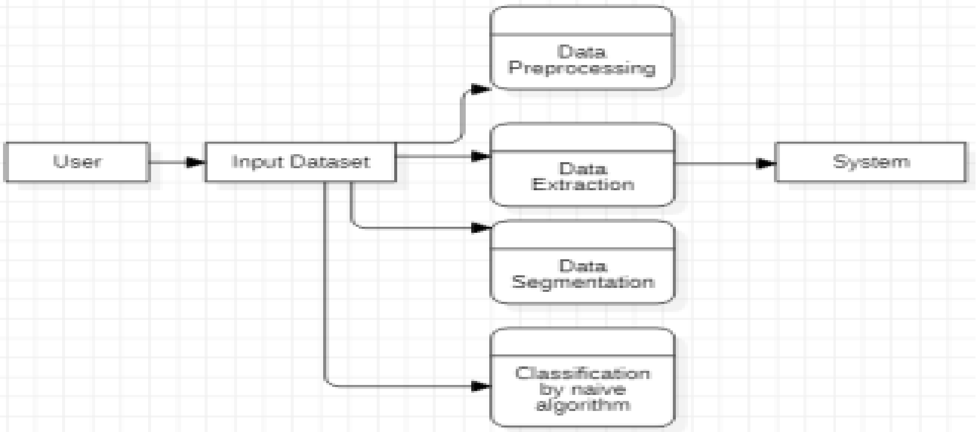


Figure 1: Flow of data for the prediction of news

Evaluation Metrics:

Confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known.

- True Positive (TP): when predicted fake news pieces are annotated as fake news.
- True Negative (TN): when predicted true news pieces are annotated as true news.
- False Negative (FN): when predicted true news pieces are annotated as fake news.
- False Positive (FP): when predicted fake news pieces are annotated as true news.

Recall - Out of all the positive classes, how much we predicted correctly.

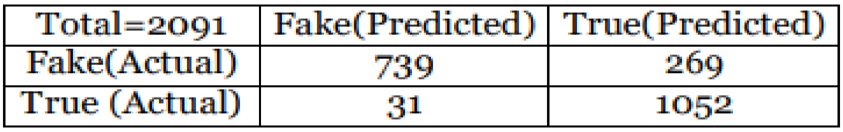
Precision -Out of all the classes, how much we predicted correctly.

F-measure - F-score helps to measure Recall and Precision at the same time.

IV. RESULTS AND DISCUSSION

Confusion Matrix: After applying various extracted features (Post-tagging and Tf-Idf) on naïve bayes classifier, the confusion matrix showing actual set and predicted sets are mentioned below:

Table 1. Confusion Matrix



Total=2091	Fake(Predicted)	True(Predicted)
Fake(Actual)	739	269
True (Actual)	31	1052

Evaluation of Classifier: Evaluation of machine learning algorithm using various metrics like :

1.Accuracy : TN +TP /TN +FP +TP +FN

2.Precision: TP/ TP +FP

3.Recall: TP/ TP +FN

4.F1-Score: = 2* Precision∗Recall /Precision+Recall

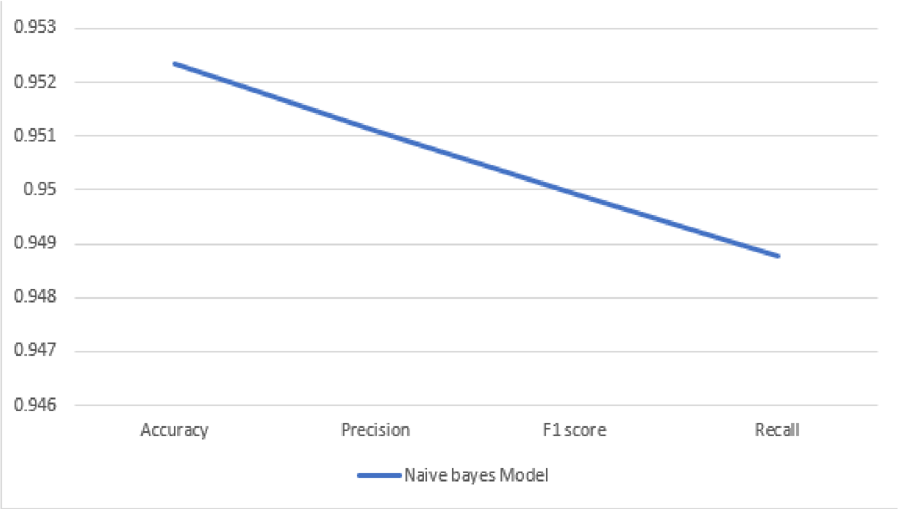


Figure 2: Evaluation of Classifier

V. CONCLUSION

The task of classifying news manually needs in-depth data of the domain and experience to spot anomalies within the text. During this analysis, we tend to mention the matter of classifying faux news articles victimization by machine learning model and NLP techniques. the info we tend to utilize in our work is Kaggle dataset and contains news articles folks Election results. the first aim is to spot patterns in text that differentiate faux articles from true news. We tend to extract totally different matter options through NLP techniques associated used for feature set as an input to the models. The Multinominal Naïve Bayes classifier was trained and parameter-tuned to get optimum accuracy. we tend to use multiple performance metrics to induce the accuracy. The accuracy of naïve Bayes classifier 95%.

VI. REFERENCES

[1] M. Granik and V. Mesyura, “Fake news detection using naive bayes classifier,” in 2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON). IEEE, 2017, pp. 900–903.

[2] U. Sharma, S. Saran, and S. M. Patil, “Fake news detection using machine learning algorithms,” International Journal of Creative Research Thoughts (IJCRT), vol. 8, no. 6, 2020.

[3] P. L. S. VU1F1819026, “Fake news and message detection,” Ph.D. dissertation, UNIVERSITY OF MUMBAI.

[4] S. A. García, G. G. García, M. S. Prieto, A. J. M. Guerrero, and C. R. Jiménez, “The impact of term fake news on the scientific community scientific performance and mapping in web of science,” Social Sciences, vol. 9, no. 5, 2020.

[5] A. D. Holan, 2016 Lie of the Year: Fake News, Politifact, Washington, DC, USA, 2016.

[6] S. Kogan, T. J. Moskowitz, and M. Niessner, “Fake News: Evidence from Financial Markets,” 2019.

[7] A. Robb, “Anatomy of a fake news scandal,” Rolling Stone, vol. 1301, pp. 28–33, 2017.