International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 03 | Mar 2021 www.irjet.net p-ISSN: 2395-0072

Machine Learning and NLP based Fake News Detection

Ketan Pande[1], Rahul Prajapati[1], Sadik Pathan[1], Akshay Patil[1]

[1]Department of Computer Engineering, Wagholi, Pune

Abstract: Fake news has been a problem ever since the internet boomed. The easier access and exponential growth of the knowledge offered on social media networks has created it knotty to largely differentiate between false and true information. Opposing such fake news is important because the world's view and mind set is shaped by information. People form their own opinions through the day-to-day news. If this information is false it can have devastating consequences. The quality of social media networks is additionally at stake wherever the spreading of pretend data is prevailing. Machine learning and Natural Language Processing has competed a significant role in classification of the data though with some limitations. The need of an hour is to stop these types of fake news especially in the developing countries like India, and focus on the correct, proper news article which will not affect people mentality negatively.

Keywords: Machine Learning, Scikit-learn, supervised learning, NlTK, tfidf, NLP, Flask, bagging, boosting, etc.

I- INTRODUCTION

In the today's world most of the information are hazy available on various platforms such as twitter, blog, online e-paper, social media.

Today's youngsters spent their most of the time on social media or internet. News on social media is additional appealing and fewer expensive compared to optional ancient news organisation and it's simple to do that 3 magical things share, like and comment however despite giving the profit, this category of reports from social media is minor than another earliest news sources. In Today's world, anybody can post the content over the internet and some individuals hence posts some misleading information. Falsified news is any textual or non-textual content that is fake News and is generated so the readers will start believing in something which is not true. Opposing this fake news is prime thing because the world's view is shaped by information. Many Machine Learning Algorithm have been used on different types of datasets to classify the news is fake or real.

The main objective is to discover the faux news, which is a classic text classification drawback with an undemanding proposition. The projected system helps to seek out the authenticity of the news. If the news isn't real, then the user is suggested with the relevant article.

II- LITERATURE SURVEY

[1] Fake news detection using machine learning approaches proposed neural networks and convolutional neural network to find out given news is fake or real. The major problem in this research paper is cannot handle imbalance data, it gives underflow and overflow problem, due to that effect on performance and accuracy. Ever-changing characteristics of news pose the challenge in categorization of fake and real news

[2] Quick technological advancement have approved newspapers and journalism to be distributed over the online and the rise of Twitter, YouTube, Instagram, Facebook and a few other social networking sites. Networking Siteshave become astimulating methodology to talk for individuals with each other and provide schemes and thoughts. Critical components of an individual these networking sites are fast sharing of knowledge. Specifically during this state of affairs, exactness of the news or info distributed is essential.

Fake news spreading on completely different networking sites has become the fore most regarding issue. Pretend news has majorly influenced everyday lives and also the social requests of the many individuals & caused some negative impacts. Here, the most thorough electronic databases are broken down to take a larger look into articles regarding identification of news that's pretend on networking sites mistreatment associate degree economical practice of literature review.

The elemental purpose to study this can be revealing the benefits that AI uses for the knowledge regarding pretend news & its ending in one application or the opposite. Consequently, assumptions were created that the victory of processed reasoning gadgets is over 90%. This can be accepted to be a manual for anyone associated with this field (researchers and individuals).

III – BACKGROUND

1. Machine Learning

Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for building mathematical models and making predictions using historical data or information. Currently, it is being used for recognition, speech- recognition, email

filtering, Facebook- tagging, recommender system, and many more.

I. Supervised Learning:

Supervised learning is the type of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).

TYPES OF SUPERVISED LEARNING:

1. Random Forest:

Random Forest is a popular type Supervised Machine Learning algorithm. It can be used as both Classification and Regression in ML. It is solely based on ensemble learning, which is a process of combining the multiple models to solve difficult problems.

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

Ex. Banking, Medicine, Marketing etc.

2. Naïve Bayes:

Naïve Bayes is also a supervised Machine Learning algorithm, which is based on bayes theorem and is used for solving classification problems.

It is mostly used in text classification which includes a high-dimensional training dataset. Naive Bayes classifier is one of the simple, most effective and probabilistic classification algorithms which predicts in the basis of probability of an object.

Ex: Spam Filtration, Sentiment analysis etc.

3. Decision Tree:

Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

3. Logistic Regression:

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression can be used to classify the observations using various types of data and can easily determine the most effective variables used for the classification.

2. Natural Language Processing (NLP):

NLP stands for Natural Language Processing, which is a part of Computer Science, Human language, and Artificial Intelligence. It is the technology that is used by machines to understand, analyse, manipulate, and interpret human's languages. It helps developers to organize knowledge for performing tasks.

Ex. Translation, automatic summarization, Named Entity Recognition (NER), speech recognition, relationship extraction, and topic segmentation

3. TFIDF:

TF-IDF is an information retrieval and information extraction subtask which basically aims to express the importance of a word to a document which is part of a collection of documents which we usually name a corpus. It is usually used by some search engines to help them obtain better results which are more relevant to a specific query.

4. NLTK:

NLTK (Natural Language Toolkit) is a suite that contains libraries and programs for statistical language processing. It is one of the most powerful NLP libraries, which contains packages to make machines understand human language and reply to it with an appropriate response.

5. Evaluation metrics:

Evaluation metrics explain the performance of a model. An important aspect of evaluation metrics is their capability to discriminate among model results.

I. Confusion matrix

The confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data. It can only be determined if the true values for test data are known. Confusion matrix can use some of following parameter to calculate performance of binary model. For binary classification problem, we should have use 2 * 2 matrix.

Target variable have two values:

1. Positive
2. Negative

PREDICTED VALUES	ACTUAL VALUES: POSITIVE	ACTUAL VALUES: NEGATIVE
POSITIVE	560	60
NEGATIVE	50	330

TERMINOLOGIES OF CONFUSION MATRIX

i. True Positive (TP):

In True positive, Model has given prediction TRUE, and the real or actual value was also TRUE**.**

2. True Negative (TF):

In True Negative, Model has given prediction FALSE, and the real or actual value was also FALSE.

3. False Positive (FP):

In False Positive**,** Model has predicted TRUE, but the actual value was FALSE. It is also called a Type-I error.

4. False Negative (FN):

In False Negative, Model has predicted FALSE, but the actual value was TRUE, it is also called as Type-II error**.**

6. Classification Report

A Classification repor**t** is used to measure the quality of predictions from a classification algorithm.

TERMS OF CLASSIFICATION REPORT

1. Accuracy:

It defines how often the model predicts the correct output.

2. Precision:

It can be defined as the number of correct outputs provided by the model or out of all positive classes that have predicted correctly by the model, how many of them were actually true.

3. Recall:

It is defined as the out of total positive classes, how our model predicted correctly. The recall must be as high as possible.

4. F1-Score:

F1 Score is the weighted average of Precision and Recall.

IV- PROPOSED SYSTEM

1. System Architecture

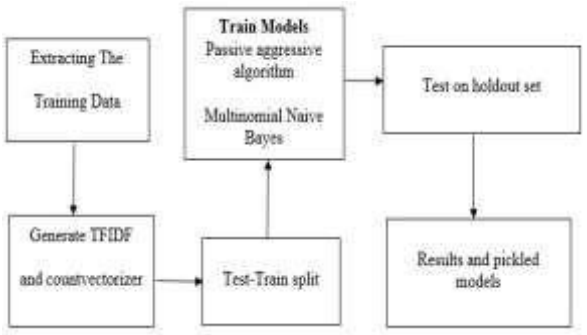

Figure: architecture of Fake news Detection

V- CONCLUSIONS

The concept of deception detection in social media is particularly new and there is ongoing research in hopes that scholars can find more accurate ways to detect false information in this booming, fake-news-infested domain. For this reason, this research may be used to help other researchers discover which combination of methods should be used in order to accurately detect fake news in social media.

It is important that we have some mechanism for detecting fake news, or at the very least, an awareness

that not everything we read on social media may be true, so we always need to be thinking critically. This way we can help people make more informed decisions and they will not be fooled into thinking what others want to manipulate them into believing.

VI- BIBLIOGRAPHY

[1] A. N. K. Movanita, "BIN: 60 Persen Konten Media Sosial adalah Informasi Hoaks (BIN: 60 percent of social media content is hoax)," 2018. [Online]. Available: https://nasional.kompas.com/read/2018/03/15/0647555 1/bin-60- persen-konten-media- social-adalah-informasi-hoaks

[2] S. Kumar, R. Asthana, S. Upadhyay, N. Upreti, and M. Akbar, "Fake news detection using deep learning models: A novel approach," Transactions on Emerging Telecommunications Technologies, 2019.

[3] F. A. Ozbay and B. Alatas, "Fake news detection within online social media using supervised artificial intelligence algorithms," Physica A: Statistical Mechanics and its Applications, vol. 540, p. 123174, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/p ii/S0378437119317546

[4] D. Pomerleau and D. Rao, "Fake News Challenge," 2017. [Online]. Available: https://www.fakenewschallenge.org

[5] A. Thota, "Fake News Detection: Deep Learning approach," SMU Data Science Review, vol. 1, no. 3, pp. 1–20, 2018.

[6] T. Saikh, A. Anand, A. Ekbal, and P. Bhattacharyya, "A Novel Approach Towards

Fake News Detection: Deep Learning Augmented with Textual Entailment Features," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11608 LNCS, pp. 345– 358, 2019.

[7] H. S. Nugraha and S. Suyanto, "Typographic-Based Data Augmentation to Improve a Question Retrieval in Short Dialogue System," in 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), dec 2019, pp. 44–49. [Online]. Available: https://ieeexplore.ieee.org/document/9034594 [8] T. Library, "Characteristics of Fake News & Media Bias," [Online]. Available: https://libguides.tru.ca/fakenews/characteristic s.

[9] N. S. SRIJAN KUMAR, "False Information on Web and Social Media: A Survey," p. 35, 2018. [4] K. shu, "Fake News Detection on Social Media: A Data Mining perspective," p. 15, 2017.

[10] The New York Times, "As Fake News Spreads Lies, More Readers Shrug at the Truth," [Online]. Available: https://www.nytimes.com/2016/12/06/us/faken ews-partisan-republican-democrat.html. [Accessed 14 04 2018].