The purpose of this project is to work on the fake news dataset and to detect the fake news. This project gives an overview of some of the most popular machine learning algorithms- Naive Bayes, SVM, K-NN and Logistic Regression. The news is categorized as fake, real, satire, bias or hate based on two different datasets. The term ‘Fake News’ is used here to define the unlike problems from fake news to satire news or the hate news. Online fake news has been a topic of interest these days and has been used in multitude of ways. I will be using machine learning algorithms in Python with Scikit-Learn to train and test the data to find out the accuracies of each algorithm. Also, the description of the algorithms is presented and comparison of their performance to find out which algorithm is suitable for this kind of text dataset.

Keywords: *dataset, Machine Learning-Regression methods, mean absolute error, R2-score*

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ML	Machine Learning
AI	Artificial Intelligent
DA	Data Analytic
MSE	Mean Squared Error
RMSE	Root Mean Squared Error

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Social media for news consumption is a double-edged sword. On the one hand, its low cost, easy access, and rapid dissemination of information lead people to seek out and consume news from social media.On the other hand, it enables the wide spread of “fake news”, i.e., low quality news with intentionally false information. The extensive spread of fake news has the potential for extremely negative impacts on individuals and society. Therefore, fake news detection on social media has recently become an emerging research that is attracting tremendous attention.

Fake news detection on social media presents unique characteristics and challenges that make existing detection algorithms from traditional news media ineffective or not applicable. First, fake news is intentionally written to mislead readers to believe false information, which makes it difficult and nontrivial to detect based on news content; therefore, we need to include auxiliary information, such as user social engagements on social media, to help make a determination. Second, exploiting this auxiliary information is challenging in and of itself as users’ social engagements with fake news produce data that is big, incomplete, unstructured, and noisy.

1.1 DOMAIN OVERVIEW:

Machine learning is to predict the future from past data. Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using python. Process of training and prediction involves use of specialized algorithms. It feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. Machine learning can be roughly separated in to three categories. There are supervised learning, unsupervised learning and reinforcement learning. Supervised learning program is both given the input data and the corresponding labelling to learn data has to be labelled by a human being beforehand. Unsupervised learning is no labels. It provided to the learning algorithm. This algorithm has to figure out the clustering of the input data. Finally, Reinforcement learning dynamically interacts with its environment and it receives positive or negative feedback to improve its performance.

Data scientists use many different kinds of machine learning algorithms to discover patterns in python that lead to actionable insights. At a high level, these different algorithms can be classified into two groups based

on the way they “learn” about data to make predictions: supervised and unsupervised learning. Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modelling is the task of approximating a mapping function from input variables(X) to discrete output variables(y). In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be

bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc.

Fig 1.1 Process of Machine learning

Supervised Machine Learning is the majority of practical machine learning uses supervised learning. Supervised learning is where have input variables (X) and an output variable (y) and use an algorithm to learn the mapping function from the input to the output is y = f(X). The goal is to approximate the mapping function so well that when you have new input data (X) that you can predict the output variables (y) for that data. Techniques of Supervised Machine Learning algorithms include logistic regression, multi-class

classification, Decision Trees and support vector machines etc. Supervised learning requires that the data used to train the algorithm is already labelled with correct answers. Supervised learning problems can be further grouped into Classification problems. This problem has as goal the construction of a succinct model that can predict the value of the dependent attribute from the attribute variables. The difference between the two tasks is the fact that the dependent attribute is numerical for categorical for classification. A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes. A classification problem is when the output variable is a category, such as “red” or “blue”.

There exists a large body of research on the topic of machine learning methods for deception detection, most of it has been focusing on classifying onlinreviews and publicly available social media posts. Particularly since late 2016 during the American Presidential election, the question of determining 'fake news' has also been the subject of particular attention within the literature.

Conroy, Rubin, and Chen outlines several approaches that seem promising towards the aim of perfectly classify the misleading articles. They note that simple content-related n-grams and shallow parts-of-speech tagging have proven insufficient for the classification task, often failing to account for important context information. Rather, these methods have been shown useful only in tandem with more complex methods of analysis. Deep Syntax analysis using Probabilistic Context Free Grammars have been shown to be particularly valuable in combination with n-gram methods. Feng, Banerjee, and Choi are able to achieve 85%-91% accuracy in deception related classification tasks using online review corpora.

CHAPTER 2

LITERATURE SURVEY

2.1 GENERAL

A literature review is a body of text that aims to review the critical points of current knowledge on and/or methodological approaches to a particular topic. It is secondary sources and discuss published information in a particular subject area and sometimes information in a particular subject area within a certain time period. Its ultimate goal is to bring the reader up to date with current literature on a topic and forms the basis for another goal, such as future research that may be needed in the area and precedes a research proposal and may be just a simple summary of sources. Usually, it has an organizational pattern and combines both summary and synthesis.

A summary is a recap of important information about the source, but a synthesis is a re-organization, reshuffling of information. It might give a new interpretation of old material or combine new with old interpretations or it might trace the intellectual progression of the field, including major debates. Depending on the situation, the literature review may evaluate the sources and advise the reader on the most pertinent or relevant of them.

2.2 REVIEW OF LITERATURE SURVEY:

Title: The Definitional Challenges of Fake News

Author: Haiden,L.., & Althuis, j

Year : (2018).

Description:

Rapid determination of soil organic matter (SOM) using regression models based on soil reflectance spectral data serves an important function in precision agriculture. “Deviation of arch” (DOA)-based regression and partial least squares regression (PLSR) are two modelling approaches to predict SOM. However, few studies have explored the accuracy of the DOA

based regression and PLSR models. Therefore, the DOA-based regression and PLSR were applied to the visible near-infrared (VNIR) spectra to estimate SOM content in the case of various dataset divisions. A two-fold cross-validation scheme was adopted and repeated 10000 times for rigorous evaluation of the DOA-based models in comparison with the widely used PLSR model. Soil samples were collected for SOM analysis in the coastal area of northern Jiangsu Province, China. The results indicated that both modelling methods provided reasonable estimation of SOM, with PLSR outperforming DOA-based regression in general. However, the performance of PLSR for the validation dataset decreased more noticeably. Among the

four DOA-based regression models, a linear model provided the best estimation of SOM and a cut-off of SOM content (19.76 g kg−1), and the performance for calibration and validation datasets was consistent. As the SOM content exceeded 19.76 g kg−1, SOM became more effective in masking the spectral features of other soil properties to a certain extent. This work confirmed that reflectance spectroscopy combined with PLSR could serve as a non-destructive and cost-efficient way for rapid determination of SOM when

hyper spectral data were available. The DOA-based model, which requires only 3 bands in the visible spectra, also provided SOM estimation with acceptable accuracy.

Title: Media-Rich Fake News Detection: A Survey

Author: Shivam B. Parikh and Pradeep K. Atrey /2018

Year: 2018

Description:

Fake News has been around for decades and with the advent of social media and modern day journalism at its peak, detection of media-rich fake news has been a popular topic in the research community. Given the challenges associated with detecting fake news research problem,researchers around the globe are trying to understand the basic characteristics of the problem statement.

Linguistic Features based Methods, Deception Modeling based Methods, Clustering based Methods, Predictive Modeling based Methods, Content Cues based Methods, Non-Text Cues based Methods**.**

Although this form of method is often deemed to be better than cue-based methods it unfortunately still does not extract and fully exploit the rich semantic and syntactic information in the content.

Title: Automatic Deception Detection: Methods for Finding Fake News

Author: Niall J. Conroy, Victoria L. Rubin, and Yimin Chen

Year: 2015

Description:

This research surveys the current state-of-the-art technologies that are instrumental in the adoption and development of fake news detection. “Fake

have significant negative societal effects Common assimilation methods face some difficulties due to the scarce, constant, or similar nature of the input parameters. For example, yield spatial heterogeneity simulation, coexistence of common assimilation methods and the nutrient module, and time cost are relatively important limiting factors. To address the yield simulation problems at field scale, a simple yet effective method with fast algorithms is presented for assimilating the time-series HJ-1 A/B data into the WOFOST model in order to improve the spring maize yield simulation. First, the WOFOST model is calibrated and validated to obtain the precise mean yield. Second, the time-series leaf area index (LAI) is calculated from the HJ data using an empirical regression model. Third, some fast algorithms are developed to complete assimilation. Existing datasets have some limitations that we try to address in our data repository. For example, Buzz Feed Newsonly contains headlines and text for each news piece and covers news articles from very few news agencies.

Title: Automatic Online Fake News Detection Combining Content and Social Signals

Author: Della Vedova, M. L., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M., & de Alfaro, L**.**

Year: 2018

Description:

The proliferation and rapid diffusion of fake news on the Internet highlight the need of automatic hoax detectionsystems. In the context of social networks, machine learning (ML) methods can be used for this purpose

This empirical regression model that was developed in Kansas was successfully applied directly in Ukraine. The model forecast winter wheat production in Ukraine six weeks prior to harvest with a 10% error of the official production numbers. In 2009 the model was run in real-time in Ukraine and forecast production within 7% of the official statistics which were released after the harvest. The same regression model forecast winter wheat production in Ukraine within 10% of the official reported production numbers six weeks prior to harvest. Using new data from MODIS, this method is simple, has limited data requirements, and can provide an indication of winter wheat production shortfalls and surplus prior to harvest in regions where minimal ground data is available. We then tested the chatbot with a completely independent

set of news, whose content was not part of the three previously mentioned datasets