***Classifying fake news articles using natural language processing to identify in article as a supervised lear***

***Abstract*** *-* The focus of the problem is to design the data science tools using various data related to real and fake news. Machine learning capability will automatically upgrade itself when there is fake news detected. Designing a flawless machine learning through data science has been done is the project.***LSTM*** networks are very good at holding long term memories or in other words, the prediction of nth sample in sequence of test samples can be influenced by an input that was given many times steps before. The long short type memory may or may not be retained by the network depending upon the data. Sherstinsky (2020) has said that long term dependencies of the network are processed by its Gating mechanisms. The network can store or release memory on the go through the gating mechanism. Thus LSTM is a good choice for such sequences which have long term dependencies in it. Therefore LSTM is used over other existing models

***Existing system :***

Detecting fake news is believed to be a comple x task and much harder than detecting fake product reviews.The open nature ofithe web and soci al media, inaddition to the recent advance incompu ter technologies, simplifies the process ofcreating and spreading fake news. While it’s easier to unde rstand and trace the intention and the impact ofifake reviews, the intention and the impact oficreating propaganda by spreading fake newscannot be meas ured or understood easily. For instance, it is clear that fake review affects the product owner, customers, and online stores; on th e other hand, it is not easy to identify the entitiesaffected by the fake news. This is because identifying these entities requires measuring the news propagation, which has shownto be complex and resource intensive..

**Working of Existing System:**

Each is a representation ofiinaccurate or deceptive reporting. Furthermore, the authors weight the different kinds ofifake news and the pros and cons of using different text analytics and predictive mod‐el ling methods in detecting them. In their paper, the y separated the fake news types into 3 groups:- 1. Serious fabrications are news not published in m ainstream or participant media, yellow press, ori ta bloids, which, as such, will be harderito collect [3]. 2. Large‐Scale hoaxes are creative and unique and often appear on multiple platforms. The authors argued that it may require methods beyond text anal ytics to detect this type ofifake news. 3.Humorous fake news is intended by their writersto be entertaining, mocking, and even absurd. Acc ording to the authors, the nature ofithe style ofthis type ofifake news could have an adverse effect on the effectiveness ofitext classification techniques. It starts with preprocessing the dataset by removin g unnecessary characters and words from the data.The n‐ gram features are extracted, and a matrix offeatures is formed representing the documents inv olved. The last step in the classification process is to train the classifier. We investigated different classifiers to predict thei class ofithe documents. We specifically investigated 6 different machine learning algorithms, namely, stochastic gradient descent(SGD), SVM, linear support vector machines (LS VM), K‐nearest neighbor (KNN), LR, and decisio n trees (DT).

Term Frequency is a method that uses word count from texts to find similarities between texts[5]. Ea ch document is represented by a vector ofiequal le ngth that contains word counts. Next, each vectoris made in such a way that the sum ofiits element s will be added to the other. Each number ofiword s is converted into opportunities forisuch a word that is present in the documents. Foriexample, ifitheiword is something document, will be represented as 1, and ifiany not in the document, itwill be set t o 0. So, each the document is represented by group s ofinames. The typical TF ofithe word w in terms ofidocument d is defined as follows: Standard Ti me = Value for Documentary / Total Number ofiD ocumentary Opposition (IDF) term w in reference to document corpus D, definedas IDF(w) D[5], by logarithm of the total number ofidocuments in the corpus divided by the number ofiletters in which t he particular name appears, and is calculated as fo llows:

Inverted document TF = 1+log (total documents /no ofidocuments with particulariterm)

TF‐IDF is a weighting metric often used in inform ation retrieval and NLP[3]. It is a statistical metric used to measure how important a term is to a docu ment in a dataset. Around 80% ofithe dataset is us ed for training and 20% for testing. After extractin g the features using either TF oriTF‐IDF, we train a machine learning classifier to decide whetheria s ample's content is truthful or fake.

**Naïve Bayes Model:**

Among the fields, that are present in the dataset, only few ofithem were used. They are link to the Facebook post with the text ofithe news article a nd the label ofithe text. Text ofithe news articles was retrieved using Facebook API[8]. News articles with labels “mixture ofitrue and false” and “no factual content” were not considered. Couple ofithe articles in the datas et are broken they do not contain any text at all ( or contain “null” as a text). These articles were ig nored as well. After such filtering data set with 1 771 news articles was obtained. The dataset was randomly shuffled, and after that divided into three subsets: training dataset, valida tion dataset, test dataset. Training dataset was use d for training the naive Bayes classifier[8]. Valid ation dataset was used for tuning some global par‐ ameters ofithe classifier. Test dataset was used t oget the unbiased estimation of how well the classifier performs on new data (it is a well known f act, that it is not correct to only have training and test datasets when parameter tuning is perfo-med, because received results on test set will be biased in this case).

For the unconditional probability ofithe fact, that a ny news article is correct all ofithe values from int erval [0.2; 0.75] with step 0.01 were considered. F or the true probability threshold all ofithe values fr om interval [0.5; 0.9] with the same step were con sidered. The best results on the validation dataset were received with the unconditional probabilityofi the fact, that any news article is correct being equ al to 0.59 and the true probability threshold being equal to 0.8.

The global parameters, that were tuned, are the unc‐ onditional probability ofithe fact, that any news ar ticle is correct and the true probability threshold.

The true probability threshold is such a value, that every article with probability to be true news articl e bigger than the threshold would be considered by the classifier as a true news article, and all otheri a rticles – as a false news article.

Consider the classification procedure ofithe naive Bayes classifier. When iterating through the wordsofithe news article that is being classified, a corner case is possible: some specific word might not be present in thei training dataset at all. For all such words it was decided to define the probability ofi t he news article being fake given that it contains th is word as 0.5. Equation (4) won’t be affected in s uch case: indeed, both nominator and denominator get multiplied by 0.5. Basically, current implementation just ignores such words.

If all of the words in the news article are unknown to the classifier (never occurred in the training da taset), the classifier reports, that it can not classify given news article

If some word occurred in the news article several times, it contributed to the total probability of the f act, that a news article is fake exactly the same number of times.

Equation (4) is computationally unstable ificalculat ed directly. This is caused by the fact, that lots ofi probabilities get multiplied, and the result ofisuch multiplication becomes close to zero really fast. Most of programming languages do not provide th e needed degree of precision, and that’s why they interpret the result of multiplication as exactlyzero [8]. Let p be the probability of the fact, that given news article is fake. One can calculate the value 1/p‐1 instead, and after that receive the value of p q uite easily. The following equation holds:

## Project Aims

The focus of the problem is to design the data science tools using various data related to real and fake news. Machine learning capability will automatically upgrade itself when there is fake news detected. Designing a flawless machine learning through data science has been done is the project.***LSTM*** networks are very good at holding long term memories or in other words, the prediction of nth sample in sequence of test samples can be influenced by an input that was given many times steps before. The long short type memory may or may not be retained by the network depending upon the data. Sherstinsky (2020) has said that long term dependencies of the network are processed by its Gating mechanisms. The network can store or release memory on the go through the gating mechanism. Thus LSTM is a good choice for such sequences which have long term dependencies in it. Therefore LSTM is used over other existing models..

## Project Objectives

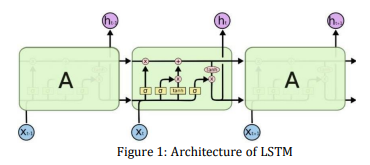
To understand if a data science application uses proper sets of data to analyse and gain information for analysis . To analyse the data and make decisions depending on the data. To design various tools related to data science according to the requirements of the project

**Proposed System :**

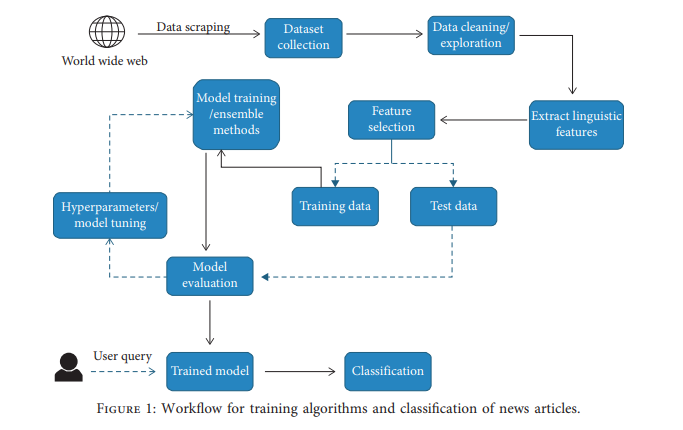
The focus of the problem is to design the data science tools using various data related to real and fake news. Machine learning capability will automatically upgrade itself when there is fake news detected. Designing a flawless machine learning through data science has been done is the project.***LSTM*** networks are very good at holding long term memories or in other words, the prediction of nth sample in sequence of test samples can be influenced by an input that was given many times steps before. The long short type memory may or may not be retained by the network depending upon the data. Sherstinsky (2020) has said that long term dependencies of the network are processed by its Gating mechanisms. The network can store or release memory on the go through the gating mechanism. Thus LSTM is a good choice for such sequences which have long term dependencies in it. Therefore LSTM is used over other existing models

**LSTM model:**

Long shortterm memory (LSTM) units are a building block f or the layers of a recurrent neural network (RNN). A LSTM unit is composed of a cell, an input gate , an output gate and a forget gate [12]. The cell is responsible for "remembering" values over a vast time interval so that the relation of the word in the starting of the text can influence the output of the word later in the sentence. Traditional neural networks cannot remember or keep the record of what all is passed before they are executed this stops the desired influence of words that comes in the sentence before to have any influence on the ending wo rds, and it seems like a major shortcoming



**System architecture :**



**Module Implementation:**

**Algorithms used :**

**System requirement specification**

**Hardware requirement specification :**

**Processor :** PIV and more

**HDD: :** 50gb and more

**Ram :** 512mb ram

**Input devices :** keyboard, mouse

**Output devices :** monitor

**Speed :** 1.2ghz

**Software requirement specification :**

**Operating system :** any os

**Programming language :** python 3.7.0 and more

**Front end :** TKinter

**Back end :** dataset

**Dataset :** leafes dataaset

**Tools :** pycham