FAKE NEWS DETECTION USING MACHINE LEARNING

[1]Dr.K.Seshadri Ramana [2]A.sai Shivani Reddy, [3]A. Sruthi [4],A.Asfiya Simran,[5]B.Iswarya

[1]Guide Profesor, [2,3,4,5,] U.G SCHOLAR
[1,2,3,4,5]Computer science and engineering.
Ravindra college of engineering for women

Email :[2]alasaishivanireddy.503@gmail.com,
[3]a.sruthi0902@gmail.com,
[4]Asfiyasimran86@gmail.com,
[5]iswaryareddy173@gmail.com

I. ABSTRACT

Fake news is the biggest scourges in our digital world. “Fake news” has been gradually increasing in the recent years due to the fact that the large part of the proliferation in social media platforms such as Facebook and other social media which are either unable or unwilling to stop the spread as well as numerous Web sites such as the boom live which specialize in publishing fake news reports. It is no longer limited to little squabbles, fake news spreads like wildfire and impacts millions of people every day.This Project comes up with the applications of Natural Language Processing for detecting the 'fake news', which are, misleading news stories which come from non-reputable sources. Only by building a model based on a count vector (using word tallies) or a (Term Frequency Inverse Document Frequency) TF-IDF matrix, (word tallies relate to how often they’re used in dataset) can only get you so far.But the models proposed earlier do not consider the important qualities like word ordering and the context. It is highly possible that the two articles which are similar in their word count may be completely different in their meaning. The data science community has responded to this by taking actions against the problem. There is a Kaggle competition called as the “Fake News Challenge” and Face book is employing AI to filter fake news stories out of users’ feeds.The main objective of our project is to detect the fake news, which is a classic text classification problem with a straight and forward proposition. We are proposing to build a model that can differentiate between “Real” news and “Fake” news. We take a dataset from Kaggle which is the combination of real and fake news and implement that dataset using various technologies like machine learning, natural language processing and deep learning.

I. INTRODUCTION

These days the news reported fake news are creating different issues from sarcastic articles to a fabricated news and plan government propaganda in some outlets. Fake news and lack of trust in the media are currently the growing problems with huge ramifications in our society. Obviously, a purposely misleading story is known as “fake news “but lately blathering social media’s discourse is changing its definition. Some of them are now using the term to dismiss the

facts counter to their preferred viewpoints.

The term 'fake news' became common parlance for the issue, particularly to describe factually incorrect and misleading articles published mostly for the purpose of making money through page views. In this paper, it is seeker to produce a model that can accurately predict the likelihood that a given article is fake news.

Facebook has been at the epicenter of much critique following media attention. They have already implemented a feature to flag fake news on the site when a user see's it, they have also said publicly they are working on to distinguish these articles in an automated way. Certainly, it is not an easy task. However, in order to solve the problem regarding the fake news, it is necessary to have a clear understanding of what Fake News is. Later, it is needed to look into how the techniques in the fields of machine learning, natural language processing helps us to detect fake news.

Fake news detection is the most dangerous types of deception because it is recently causing a way of deceiving many people, Fake news is defined to be as the prediction of the chances of a particular news article (news report, editorial, expose, etc.) which is being intentionally deceptive. We are concerned about the fake news because of the problem of fake news detection is more challenging than detecting deceptive reviews.

1.1 MACHINE LEARNING:

Machine learning is a data analytics technique which teaches the computers to do what comes naturally to humans and animals: learning from the experience. Machine learning algorithms use computational methods to "learn" information directly from data without relying on a predetermined equation as a model. The algorithms go on adaptively improving the performance as the number of samples available for learning increases. Deep learning is a specialized form of the machine learning.

From the view of SLDA, a tweet post can also be disintegrated into a bag of topics; and the

even be inferred from word distribution in each topic. Besides, the newly emerging methods: deep learning have attracted increasing attention and have been proved to be superior in some transportation studies. For instance, deep learning architecture has been proven better than the artificial neural network in traffic flow prediction (Polson and Sokolov, 2017).

RNN can be taken as multiple copies of the same network and are able to

pass information in sequence from the previous inputs to the present task. Thus, it is a powerful model for sequential data and proves valid in long speech recognition (Graves et al., 2013). If we employ RNN for a supervised learning task in language modeling, the process can be described as using a sequence of word features to predict the manual labels such as topics, sentiment (Agarwal et al., 2011), etc. One special form of RNN: Long Short-Term Memory Network (LSTM) moves one step further which is capable of learning long-term dependencies between words within the context.

The LSTM unit in the network can remember the inputs for either long or short durations. The input information from lower layers is neither converted nor eliminated because there is no conversion from lower layers to upper layers. These unique features are valued in the applications such as speech tagging.

Applications using RNN or LSTM for classification is an attractive choice for sequence labeling, which can finish a variety of tasks in topic modeling. Given that DNN models work well in language modeling, we expect them to deliver promising results in classifying the accident-related tweets.

This is mainly because DNN consists of multiple layers or stages of nonlinear information processing which captures the inter-feature correlation. Besides, it can represent features successively by higher, more abstract layers. The deep learning methods are expected to be more effective in classifying the accident-related tweets than other methods. Two networks are tested in this study: DBN and LSTM. In addition, their effectiveness in classifying short tweet contexts will be fully discussed.

With the rise in big data, machine learning has become a key technique for solving problems in areas, such as:

- Computational finance, credit scoring and algorithmic trading
- Image processing and computer vision, face recognition, motion detection, and object detection
- Energy production, price and load forecasting
- Natural language processing, for the voice recognition applications

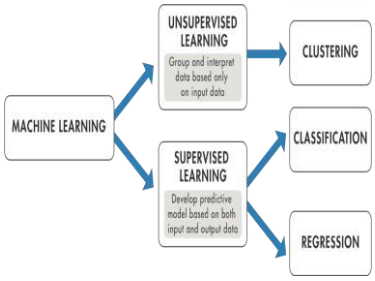
1.2 WORKING:

Machine learning uses two types of techniques such as supervised learning, which trains a model on the known input and output data so that it can predict the future outputs, and unsupervised learning, which finds hidden patterns or intrinsic

structures in the input data.

CLASSIFICATION TECHNIQUES

- Classification techniques predict the discrete responses for example, whether an email is genuine or



spam, or whether a tumor is cancerous or benign. Classification models classify input data into different categories. Typically, applications of classification include medical imaging, speech recognition, and credit scoring.

- Use classification if the data can be tagged, categorized, or separated into specific groups or classes. For example, applications for hand-writing recognition use the classification to recognize the letters and numbers. In the image processing and computer vision, the unsupervised pattern recognition techniques are used to detect the objects and image segmentation.
- Common algorithms for performing classification include support vector machine (SVM), boosted and bagged decisi on trees, k-nearest neighbor, Naïve Bayes, discriminant analysis, logistic regression, and neural networks.

Figure: Classification techniques

SUPERVISED LEARNING:

Supervised machine learning builds a model which makes the predictions based on the evidence in the presence of uncertainty. A supervised learning algorithm takes a known set of the input data and the known responses to the data and trains a model to generate reasonable predictions for the response to the new data. Supervised learning uses classification and the regression techniques to develop predictive models.

UNSUPERVISED LEARNING:

Unsupervised learning finds the hidden patterns or intrinsic structures in the given data. It is used to draw the inferences from datasets consisting of the input data without the labelled responses.

REGRESSION TECHNIQUES

- Regression techniques are used to predict the continuous responses from the data, for example, changes in temperature or fluctuations in power demand.

Typical applications for this include electricity load forecasting and algorithmic trading.

- Regression techniques are used when working with a data range or if the nature of the response is a real number, such as temperature or the time until failure for a piece of the equipment.

CLUSTERING

Clustering is one of the most common unsupervised learning techniques. It is used for exploratory data analysis to find the hidden patterns or groupings in data. Applications for the cluster analysis includes the gene sequence analysis, market research, and the object recognition. For example, if a cell phone company may want to optimize the locations where they build cell phone towers, they can use machine learning to estimate the number of clusters of people relying on their towers in a particular area.

USES OF CLUSTERING

There are unlimited reasons to cluster data. The main reason is that it allows us to build simpler, more understandable models of the world which can be acted upon easily. People naturally cluster objects for this reason all the time. Clustering algorithms automate this process and allow us to exploit the power of computer technology. A secondary use for clustering is for dimensionality reduction or data compression.

Clustering can be employed by a search engine so that the documents retrieved from the search term jaguar cluster the documents related to the jaguar animal separately from those related to the Jaguar automobile.

CATEGORIES OF CLUSTERING ALGORITHMS

Clustering algorithms can be organized by the basic approach that they employ. These approaches are also related to the type of clustering that the algorithm produces. The two main types of clustering algorithms are hierarchical and non-hierarchical.

A hierarchical clustering has multiple levels while a non-hierarchical clustering has only a single level. An example of a hierarchical clustering is taxonomy used by biologists to classify living organisms.

Figure: Clustering patterns

This algorithm randomly assigns each object to one of the k clusters and then computes the average of the points in the cluster. This cycle continues until no

further changes are made. Another way to generate non-hierarchical clustering is through density-based clustering methods, such as DBSCAN. One advantage of DBSCAN is that because it's sensitive to the density differences it can form clusters with arbitrary shapes.

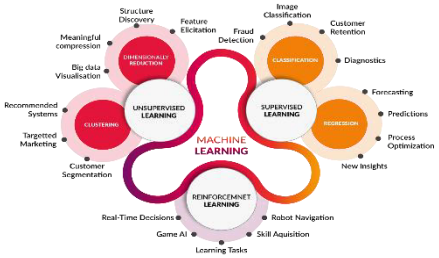
Finding the right algorithm is partly just trial and error-even highly experienced data scientists can't tell whether an algorithm will work without trying it out. Guide lines on choosing between supervised and unsupervised machine learning:

- Choose supervised learning if there is a need to train a model to make a prediction--for example, the future value of a continuous variable, such as temperature or a stock price, or a classification—for example, identify makes of cars from webcam video footage.
- Choose unsupervised learning if there is a need to explore your data and want to train the model to find a good internal representation, such as splitting data up into clusters.

Figure: Classification of Machine learning.

1.3 PROCESS OF MACHINE LEARNING

Putting all of the above observations together, we can now outline the typical process used in Machine Learning. This process is designed to maximize the chances of learning success and to effectively measure the error of the algorithm.



TRAINING:

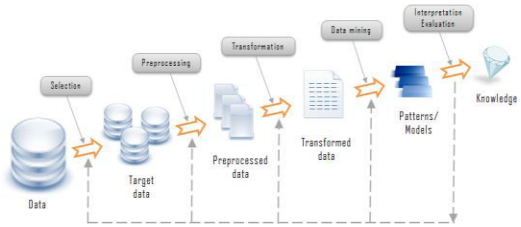
A subset of real data is provided to training model. The data includes a sufficient number of positive and negative examples to allow any potential algorithm to learn. The data scientist, experiments with a number of algorithms before deciding on those which best fit the training data.

VALIDATION:

A further subset of real data is provided by the data scientist with similar properties to the training data. This is called the validation set. The data scientist will run the chosen algorithms on the validation set and measure the error. The algorithm that produces the least error is considered to be the best.

TESTING:

It will be important to measure the mean square error of any learning algorithm that is considered implementable. The validation set should not be used to calculate this error as we have already used the validation set to choose the algorithm so that it has minimal error. Therefore, the validation set has now effectively become a part of the training set.



DATA MINING

Data mining is a process that takes data as input and outputs knowledge. it is defined as the procedure of extracting information from huge sets of data. In other words, we can say that data mining is mining knowledge or insides from data. Data pre-processing is a proven method of resolving such issues. Data processing prepares raw data for further processing. This step consists of feature selection, data cleaning and data transformation. Data gathered from different sources was consolidated, mapped and scrutinized. Some of the data that is not useful to the data mining exercise was ignored.

In this process, few attributed were transformed into required formats for example, the feature “Hour” was converted into 24-hour format. Also, the attribute values were hard-coded for better representation in the training dataset. A formal presentation of the rule and parameters of confidence, support and lift which quantify a rule is as follow.

Figure: Data mining process image

OVERVIEW OF DATAMINING TASKS

The best way to gain an understanding of data mining is to understand the types of tasks, or problems, that it can address using given dataset. At a high level, most data mining tasks can be categorized as either having to do with prediction or description. Predictive tasks allow one to predict the value of a variable based on other existing information.

1.4 CLASSIFICATION AND REGRESSION

Classification and regression tasks are predictive tasks that involve building a model to predict a target, or dependent, variable from a set of explanatory, or independent, variables. For classification tasks the target variable usually has a small number of discrete values whereas for regression tasks the target variable is continuous. Identifying fraudulent credit

card transactions is a classification task while stock price prediction is a regression task.

1.4.1 LOGISTIC REGRESSION

Logistic regression is the regression analysis and dependent upon the variables is binary numbers i.e. 0's and 1's. All regression analysis, the logistic regression is a prediction analysis. Logistic regression is used to details about data and to graphically explain the relationship between dependent binary variable and more nominal, ordinal, interval independent variables.

1.4.2 ASSOCIATION RULE ANALYSIS

Association rule analysis is a descriptive data mining task that involves associations or discovering patterns, between elements in a data set. The associations are represented in the form of rules or implications.

The most common association rule task is market basket analysis. In this case each data record corresponds to a transaction and lists the items that have been purchased as part of the transaction. It should be noted that although this is a descriptive task, highly accurate association rules can be used for predicting results.

1.4.3 CLUSTER ANALYSIS

Cluster analysis is a descriptive data mining task where the goal is to group similar objects in the same cluster and dissimilar objects in different clusters i.e cluster formation. Applications of clustering include clustering customers for the purpose of market segmentation and grouping similar documents together in response to a search engine request.

1.4.5 TERMINOLOGY AND BACKGROUND

Most prediction tasks assume that the underlying data is represented as a collection of objects or records, which, in data mining, are often referred to as instances or examples. Each example is made up of a number of variables, commonly referred to as features or attributes.

1.5 PREDICTIVE DATA MINING ALGORITHMS

We briefly describe some of the most common data mining algorithms. Because the purpose of this chapter is to provide a general description of data mining, its capabilities, and how it can be used to solve real-world problems, many of the technical details concerning the algorithms are omitted.

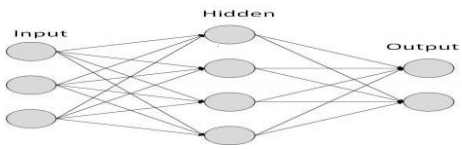
1.6 DECISION TREES

Decision tree algorithms are a very popular class of learning algorithms for classification tasks. A sample decision tree, generated from the automobile loan data. The internal nodes of the decision tree each represent a feature while the terminal nodes are labeled with a class value. Each branch is labeled with a feature value, and, when presented with an example, one follows the branches that match the attribute values for the example, until a leaf node is reached.

The class value assigned to the leaf node is then used as the predicted value for the example. In this simple example the decision tree will predict that a customer will default on their automobile loan if their credit rating is “poor” or it is not “poor” but the person is “middle aged” and their income level is “low”.

1.7 RULE-BASED CLASSIFIERS

Rule-based classifiers generate



classification rules, such as rule set. The way in which classifications are made from a rule set varies. For some rule-based systems the first rule to fire determines the classification, whereas in other cases all rules are evaluated and the final classification is made based on a voting scheme. Rule-based classifiers are very similar to decision-tree learners and have similar expressive power, computation time, and comprehensibility.

1.8 ARTIFICIAL NEURAL NETWORKS

Artificial Neural Networks (ANNs) were originally inspired by attempts to simulate some of the functions of the brain and can be used for both classification as well as regression tasks. An ANN is composed of an interconnected set of nodes that includes an input layer, zero or more hidden layers, and an output layer. The links between nodes have weights associated with them. The ANN accepts three inputs, I1, I2, and I3 and generates a single output O1.

ANNs can naturally handle regression tasks, since numerical values are passed through the nodes and are ultimately passed through to the output layer. However, ANNs can also handle classification tasks by thresholding on the output values. ANNs have a great deal of expressive power and are not subject to the same limitations as decision trees. In fact, most ANNs are universal approximators in that they can approximate any continuous function to any degree of accuracy.

Figure: A Typical Artificial Neural Network

However, this power comes at a cost. While the induced ANN can be used to quickly predict the values for unlabeled examples, training the model takes much more time than training a decision tree or rule-based learner and, perhaps most significantly, the ANN model is virtually incomprehensible and therefore cannot be used to explain or justify its predictions.

1.9 NEAREST-NEIGHBOR

Nearest-neighbor learners are very different from any of the learning methods just described in that no explicit model is ever built. That is, there is no training phase and instead all of the work associated with making the prediction is done at the time an example is presented.

Nearest-neighbor algorithms are typically used for classification tasks, although they can also be used for regression tasks. These algorithms also have a great deal of expressive power. Nearest-neighbor algorithms generate no explicit model and hence have no training time.

The nearest-neighbor method first determines the k most similar examples in the training data and then determines the prediction based on the class values associated with these k examples, where k is a user specified parameter. The simplest scheme is to predict the class value that occurs most frequently in the k examples, while more sophisticated schemes might use weighted voting, where those examples most similar to the example to be classified are more heavily weighted. People naturally use this type of technique in everyday life.

1.10 NAÏVE BAYESIAN CLASSIFIERS

Most classification tasks are not completely deterministic. That is, even with complete knowledge about an example you may not be able to correctly classify it. Rather, the relationship between an example and the class it belongs to is often probabilistic in nature. Naïve Bayesian classifiers are probabilistic classifiers that allow us to exploit statistical properties of the data in order to predict the most likely class for an example.

These methods are quite powerful, can express complex concepts, and are fast to generate and to classify new examples.

More specifically, these methods use the training data and the prior probabilities associated with each class and with each attribute value and then utilize Bayes' theorem to determine the most likely class given a set of observed attribute values. This method is naïve in that it assumes that the values for each

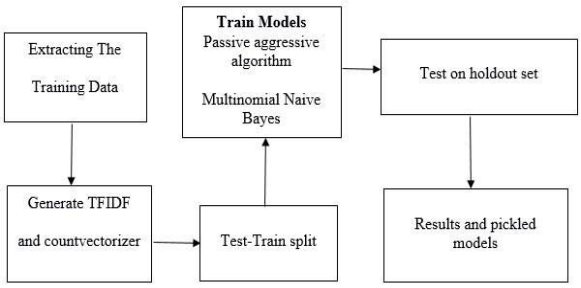
attribute are independent.

1.11 ENSEMBLE METHODS

Ensemble methods are general methods for improving the performance of predictive data mining algorithms. The most notable ensemble methods are bagging and boosting, which permit multiple models to be combined. With bagging the training data are repeatedly randomly sampled with replacement, so that each of the resulting training sets has the same number of examples as the original training data but is composed of different training examples.

TEXT CLASSIFICATION:

This is one of the important and typical task in supervised machine learning (ML). Assigning categories to documents, which can be a web page, library book, gallery etc. has many applications like e.g.,



spam filtering, sentiment analysis etc. Text classification is a smart classification of text into categorion

- Supervised Text Classification
- Unsupervised Text Classification

TF: While counting the number of words in each document, This will give more weightage to longer documents than shorter documents. To avoid this, we use frequency TF - Term Frequencies, in each document.

TF-IDF: we can even reduce the weightage of more common words like (the, is, an etc.) which occurs in all document. This is called as TF-IDF i.e Term Frequency times inverse document frequency.

II. WORKING

A model is build based on the count victimizer or a TF-IDF matrix (i.e.) word tallies relatives to how often they are used in other articles in your dataset) can help. Since this problem is a kind of text classification, implementing a Naive Bayes classifier will be best as this is standard for text-based processing.

Now the next step is to extract the most optimal features for count vectorizer or TF-IDF -victimizer, this is done by using a n-number of the most used words, and/or phrases, lower casing or not, mainly removing the stop words which are common words such as “the”, “when”, and “there” and only using those words that appear at least a given number of times in a given text dataset.

A dataset is a collection of data which mainly comprises of single statistical data matrix, database table where every row corresponds to each member in datasets and each column represents variable. The dataset list values for each variable such as title, id, author, label etc.

We commonly collect various datasets from https://www.kaggle.com.

1. Exclusively fake news article
   - Datasets contain news article are forged.
2. Exclusively real news Data
   - This dataset contains news articles which are certain.
3. Mixed Data of fake and real news
   - This dataset contains the combination of real and fake news

3.1 DATA SET SELECTION:

Data is the most import part when you work on prediction systems. It plays a very vital role your whole project i.e., your system depends on that data. So, selection of data is the first and the critical step which should be performed properly, for our project we got the data from the government website.

3.2 DATA CLEANING AND DATA TRANSFORMATION:

After we have selected the dataset. The next step is to clean the data and transform it into the desired format as it is possible the dataset we use may be of different format. It is also possible that we may use multiple datasets from different sources which may be in different file formats.

3.3 DATA PROCESSING AND ALGORITHM IMPLEMENTATION:

After the data is been cleaned and transformed it's ready to process further. After the data has been cleaned and we have taken the required constraints The output of the algorithm is in 'yes' and 'no'. It gives the error rate and the success rate. Our work clarified in the following steps as follows:

FIRST STEP:

- The first level is splitting each sentence to deal with separately.
- The second level is removing stop words and that includes identifying the useless words in each statement like (the, a, an, etc).
- The third level is stemming which every word returns to its infinitive.

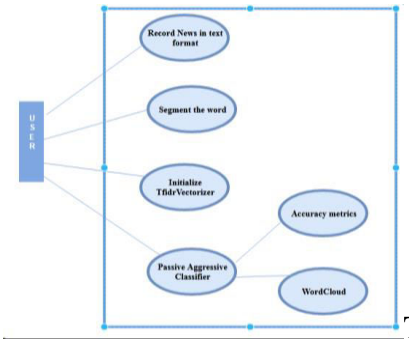
SECOND STEP:

The algorithm detects the words that appear mostly together and it shows their relationship and then this is able to predict the next word.

THIRD STEP:

Results of word embedding level will be the input to the model.

FOURTH STEP:

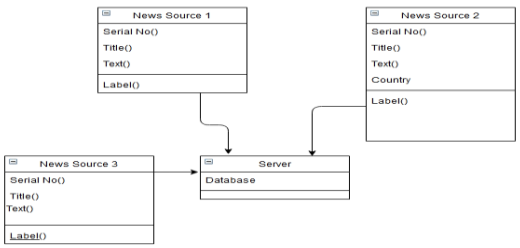


The output of step four will Getting final result determining if the piece of news is truthful or deceptive as is common in data mining problems, once the models are built, the process might be repeated with new data and new features.

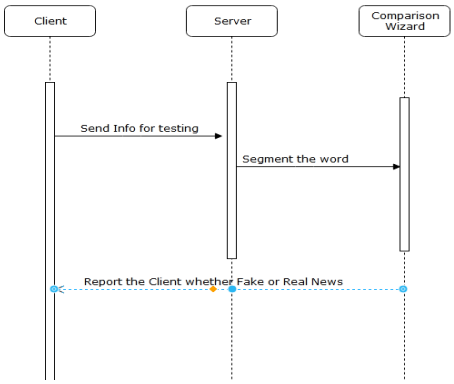
III. SYSTEM DESIGN

USE CASE DIAGRAM:

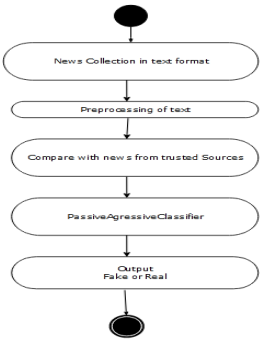
CLASS DIAGRAM:



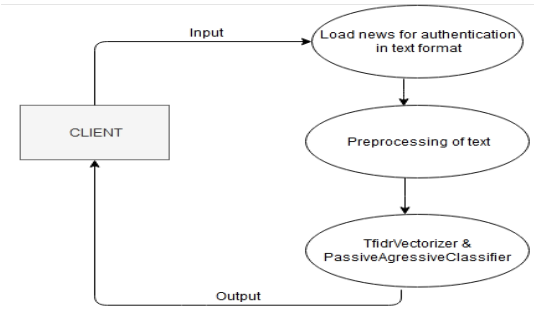
SEQUENCE DIAGRAM:



ACTIVITY DIAGRAM:



COMMUNICATION DIAGRAM:



IV. RESULT

A. Static System-

Figure 1: Static output (True)

Figure 2: Static Output

(False)

B. Dynamic System-

Figure 3: Fake News Detector (Output page)

V. CONCLUSION

With the increasing popularity of social media, more and more people consume news from social media. However, social media has also been used to spread fake news, which has strong negative impacts on individual users and society. This fake news is created by someone intentional to create wanted violence, anger in our society. Most of the times, the younger generations face the lot of effects mentally. To avoid this situation, we have created this system called fake news detector. Fake news detection system will differentiate between fake and real news from the social media.

VI. REFERENCES

[1] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, pp. 1–4, 2015.

[2] Bajaj, Samir. ""The Pope Has a New Baby!" Fake News Detection Using Deep Learning."

[3] Akshay Jain and Amey Kasbe, "Fake News Detection", IEEE International Students' Conference on Electrical, Electronics and Computer Sciences, 2018.

[4] Shivam B. Parikh and Pradeep K. Atrey, "Media-Rich Fake News Detection: A Survey", IEEE Conference on Multimedia Information Processing and Retrieval, 2018.

[5] Cade Metz. (2016, Dec. 16). The bittersweet sweepstakes to build an AI that destroys fakenews.

[6] Conroy,N.,Rubin,V.andChen,Y.(2015).Automatic deception detection: Methods for finding fake news‖ at Proceedings of the Association for Information Science and Technology,

52(1), pp.1-4.

[7] Markines,B.,Cattuto,C.,&Menczer, F. (2009, April).–Social spam detection. In Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web (pp.41-48)

[8] Rada Mihalcea , Carlo Strapparava, The lie detector: explorations in the automatic recognition of deceptive language, Proceedings of the ACL-IJCNLP