**Fake News Exposure Using Machine Learning**

**Abstract**

News is a routine in everyone's life. It helps in enhancing the knowledge on what happens around the world. Fake news is a fictional information madeup with the intension to delude and hence the knowledge acquired becomes of no use. As fake news spreads extensively it has a negative impact in the society and so fake news detection has become an emerging research area. The paper deals with a solution to fake news detection using the methods, deep learning and Natural Language Processing. The dataset is trained using deep neural network. The dataset needs to be well formatted before given to the network which is made possible using the technique of Natural Language Processing and thus predicts whether a news is fake or not. Fake news is a fallacious piece of information. Since fake news spread much faster it has become desideratum to categorize a news into real and fake. It has become a major factor in creating social disturbances and can be spread by anyone as social media is an open platform for everyone to express their ideas on any topic. Spreading of fake news can cause loss of trust in the society and also has the ability to exploit public’s thoughts into a different manner. As they are interwoven it is difficult to simply identify between real and fake news. Therefore it can be said that news act as a bonus to the society whereas deceptive news is a bane. There are many instances where fake news has caused so much of instabilities and religious clashes in the cultural aspects of the society. However the distinction between fake and real news is a difficult task as the idea of identification is full of dependencies and disparities. The large amount of information that gets spread are most of the times unverified and considered truth.In 2017, a German official stated that spreading of fake news is a huge phenomenon of dimension one could never see before. Therefore recognizing the rightness of a news is an engrossing topic and can be consummated using different philosophies.

**Existing System**

The current problem statement of this research revolves around implementing a system which would proactively measure the truthfulness of the news content that is published in the social media. The necessity of the system is to truncate the rapid spread of the falsified information that pose a threat to social media as a platform. In this project, we are striving to leverage auxiliary information available in social media to classify the news content. The main factors that are involved in the social media news ecosystem are the news articles and the users (publishers, readers etc). In this ecosystem, once the news content is published, the news is not only validated against the authenticated sources, but also the other users engagement, the publisher bias towards the published news topic and the historical credibility values of the user is taken into consideration to a more justifiable classification of the news content. On a high level, this system have computes the News-News relationship from the social media using the User-News relationship and derives the News Content Vector.

**DRAWBACKS**

* As an increasing amount of our lives is spent interacting online over the internet, more and more people tend to seek and consume news from social media, news agency homepages, search engines.
* On the other hand, it enables the proliferation of “fake news", i.e., low quality news with intentionally false information.

**Proposed System**

Most of the current research in the field of fake news detection use supervised means to classify the news as either fake or real. This involves training a classification model based on a predefined labeled dataset. Though these methods show significant results, they suffer a major drawback i.e. they require a reliably annotated dataset to train a classification model. This process is time-consuming and labor intensive as it requires careful checking of news contents as well as other additional evidence such as authoritative reports. In order to mitigate this issue, the authors of the paper propose an unsupervised learning framework, to identify the fake news. The key idea involves extracting users’ opinions on the news by exploiting the auxiliary information of the users’ engagements with the news tweets on social media and aggregating their opinions in a well-designed unsupervised way to generate the final estimation results. first extracts the users’ opinions on the news by analyzing their engagements on social media and builds a Bayesian probability graphical model capturing the complete generative process of the truths of news and the users’ opinions. Next, an efficient collapsed sampling approach is proposed to detect fake news and estimate the users’ credibility simultaneously.

**FEATURES**

The usage of the web as a medium for perceiving information is increasing daily.

The amount of information loaded in the social media at any point is enormous, posing a challenge to the validation of the truthfulness of the information

**MODULE DESCRIPTION**

**1. Data Preprocessing:**

The dataset comprised of information (in the form of JSON files) of news articles from BuzzFeed and PolitiFact. Along with this, the dataset provided information about users and user-news interaction in the form of CSV files. Furthermore, the dataset included real and fake news content. Based on our intuition and research, we found that the “body” of the news articles, i.e, main content, best represents the news articles core information and variance. After extracting the body of all the news articles we constructed a data frame of the news-ids and the body of the text. A label was also added to the data frame to indicate whether the news article is fake or real.

**2. Derive Relationships**

Once the important aspects of the data were identified for feature extraction, the next step involved establishing relationships among the news articles that can be leveraged by classification model. The relationship between the news articles was established using a graph data structure. Here, the news articles act as the nodes of the graph and the edges between the nodes represent the relationships among the new articles. The edges between the nodes of the graphs were established using the following steps.

**3. Feature Extraction**

For the fake news detection, the actual news data (body of the news article) is being considered as features. But the data is in the form of text. It is known that for the machine learning analysis, text data does not work well. So the text data has to be converted into a numerical representation. This process is called vectorization. Every record (i.e news article in this case) should be converted into a vector. There are several techniques/algorithms which can convert text to vector. Below is the list of such techniques.

**4. Training Classification Model**

As explained in the previous sections, the adjacency matrix, feature vectors and the labels form the input for the classification model. The fake news detection task has network-based input in the form of adjacency matrix representing the relationship between news articles. As explained before, this relationship is derived from the information provided in the form of users and their association with the news articles (post, share, retweet etc). As one can observe, the data has the pattern of a graph with nodes or vertices being the news articles and edges as the relationship between them. The classification task needed a model which can utilize this network or graph-based architecture of the dataset and the feature vectors generated for the news articles using BERT. Traditional machine learning models utilize just the feature vectors for prediction and will not leverage the graph-based architecture of the data. The solution to this problem was obtained through Graph Attention Networks