FAKE NEWS ANALYSIS

Chirag Patel[*1]**, Kaushal Gor**[*2]

PG Student, MCA, Parul Institute of Engineering and Technology–MCA, Parul University, Vadodara, Gujarat, India

[2]Assistant Professor, MCA, Parul Institute of Engineering and Technology–MCA, Parul University, Vadodara, Gujarat, India

ABSTRACT

We all know that the internet changed our life from the 90's onwards and the data explosion happened in the last two decades due to mobile usage and social platforms. At the same time there is a high percentage of threats evolues and we need to tackle this in a better way. That threat is "FAKE NEWS" as the name suggests it's misleading information spreading into the public domain and it's kind of causing problems. for that stance detection is used as a primary core principle in sentiment analysis.

Keywords: Research Paper, Technical Writing, Science, Engineering and Technology

I.   INTRODUCTION

1.1 What is fake news?

Any information that looks authentic and after misleads us by falsely representing the actual information or the truth.

How was it identified? Traditionally, fake news was identified using our common sense by comparing the news from trustworthy sources.

How viral fake news is generated? In the past, humans manually generated the fake news by providing false information. But from last two decades scenario has been changed. False news spreading faster then truth on social medias. The massive spread of false news has the potential to have an extremely negative effect on individuals and society. This disinformation may be spread in order to re-establish popularity or just for fun. As a result, detecting fake news on social media has recently become emerging research that has garnered a great deal of attention.

The present work on the detection of false news is mainly based on supervised methods. they aim to build a classification model considering different sets of features including news content, message propagation . though they have shown some promising result, these supervised method suffer from a critical stamming. they require a reliably meaningful pre-annotated dataset to train a classification model. To solve this problem my proposed system using a pre-text processing using stopword and porterstemmer library. Pre-processed text now vectorized using TF-IDF vectorizer. Generated features are pass to LSTM layer to Dense layer which gives final node of probability of whether the news is fake or not.

II.   Application Areas [Page Layout]

❏   Textual Dataset

A dataset is a collection of data pieces that can be treated by a computes as a single unit for analytic and prediction purpose.

❏   features

Features are individual independent variable that act like a input in your system. while making the predications, model use such features to make the predictions.

❏   models

A model represents what was learned by a machine learning algorithm. the model is the "thing" that is saved after running a machine learning algorithm on training data and represents the rules, numbers, and any other algorithm specific data structures required to make predictions.

❏   Images

News images that have been faked by various means, generally to promote an ideological agenda or to manipulate the emotions of the viewer.

III. **Methodologies [Page Style]**

Bag of words:
A bag-of-words is a model which extract features from the input text and use machine learning models to train and test for classification of input texts. it mainly used to classify documents or text or sentences.
as the mane suggests BOW represent occurrences of the words in a document and mainly depends on the known words vocabulary and occurrence of the known words.

Stemming:
Stemming is a process of removing the suffix from a word and reduce it to root word. Stemming is an important part of the pipelining process in NLP. the input to stemmer is tokenized words. Example stemming from the root term "like" includes: "likes", "liked", "likely", and "liking".

TF-IDF:
TF-IDF: TF-IDF refers to the reverse term document frequency. it's a statistical measure that evaluates how to relevant a word is to a document in a collection of documents.
it has many uses, most importantly in automated text analysis, and is very useful for scoring words in machine learning algorithms for NLP.
TF-IDF for a word in a document is calculated by multiplying two different
metrics:
Term frequency of a word in a document. there are several ways of calculating this frequency, with the simplet being a raw count of instances a word appears in a document.

$$TF = \frac{\text{no of repetitions of word in sentence}}{\text{no of words in sentence}}$$

the inverse document frequency of the word across a set of documents. this means, how common or rare a word is in the entire document. the metric can be calculated by

$$idf = \log\left(\frac{\text{no. of sentence}}{\text{no of sentence containing words}}\right)$$

Multiplying these two numbers results in the TF-IDF score of a word in a document. the higher the score, the more relevant that word is in that particular document.

IV. **Algorithm**

**Passive aggressive classifier algorithm:**

This algorithm is online leaning algorithm. Generally used for large-scale learning. passive aggressive algorithm take input in form of sequential order (vectorized data) and machine learning model is updated. this is very useful in situations where there is huge amount of data and it is computationally infeasible to train the entire dataset becaus of the size of data. This is very good algorithm for detecting fake news on social media like twitter.

Passive: If prediction is correct, keep the model and do not make any changes.

Aggressive: If the prediction is incorrect, make changes to model.

Long short-term memory network:

It is special kind of recurrent neural network that is capable of learning long term dependencies in data. this is achieved because the recurring module of the model has a combination of four layers interacting with each other.
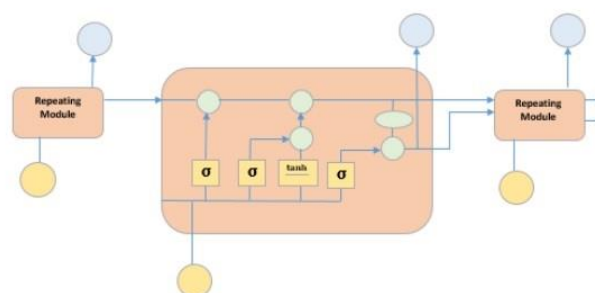
An LSTM module has a cell state and three gates which provide them with the power to selectively learn, unlearn or retain information from each of the units.

Dense layer:

Dense implements the operation: output = activation(dot(input,kernel) + bias) where activation is the element-wise activation function passed as the enable argument, the nucleus is a weight matrix created by the layer, and the bias is a bias vector created by the layer. these are all attribute of dense. Result is the output and it will be predicted news is fake if value is > 0.5 otherwise it's true.
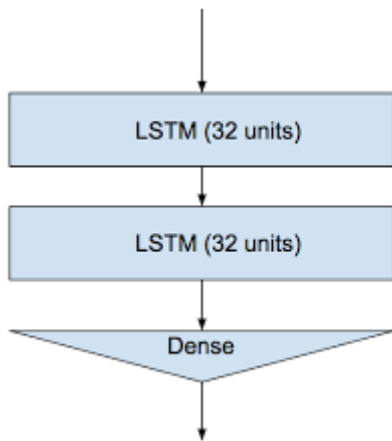
V. Tools and Technologies

❑   JUPYTERLAB

Jupyter is free, open-source, web based interactive development environment. Jupyter support a wide range of workflow in data science, scientific computing, and machine learning.

❑   MATLAB

Matlab is a proprietary multi-paradigm programming language and numeric computing environment developed by Mathworks.

❑   GOOGLE COLAB

Colab is a free jupyter notebook environment that runs entirely in the cloud. Colab supports many popular machine learning libraries which can be easily loaded

VI. Current work in the field

A wide range of tools and related work is already available using stance detection in the market. Stance dedication deals with extracting a subject's reaction to the main claim, which is the central part of the fake news assessment. As part of the fake news challenges many models were developed using neural networks using stance detection principles.

Talos research intelligence team-SLAT developed a 50/50 weighted average ensemble method which combines CNN with pre-trained vectors and gradient-boosted decision trees. The model achieved 82.02% accuracy during the FNC challenge.

In another research the model was designed using handcrafted features, and an RNN plus attention model with TF-IDF is used to extract features to represent both headlines and bodies of the m=news articles. The model provides 86.66% accuracy level

In another work, a conditioned bidirectional LSTM with global features was used. It provides a combination of global features and local word embedding features.

This article offered a method to raise awareness about possible misinformation on Twitter based on empirical evidence, Python and Excel analysis, and the Twitter API. Tweets were assessed to show feelings (positive, negative or neutral) of a tweet that was rated positive or unfavourable to COVID-19 precautions. As well, the average subjective nature of the tweets was also assessed. Overall, the document showed a clear method for identifying and labeling a mix of subjects on Twitter as erroneous information. Efforts are showing promising results on how to deal with disinformation on Twitter. Future work will focus on cross-referencing with different sources of information, additional and enhanced feelings analysis, and analysis of broader Twitter datasets. In addition, the current work only dealt with text tweets, but due to the nature of Twitter, analysis of images containing text would be beneficial as well. By using this method, one becomes able to analyze tweets and data on different.

VII. Proposed System

The proposed system for analysing fake news by combining fake news and many Techniques.
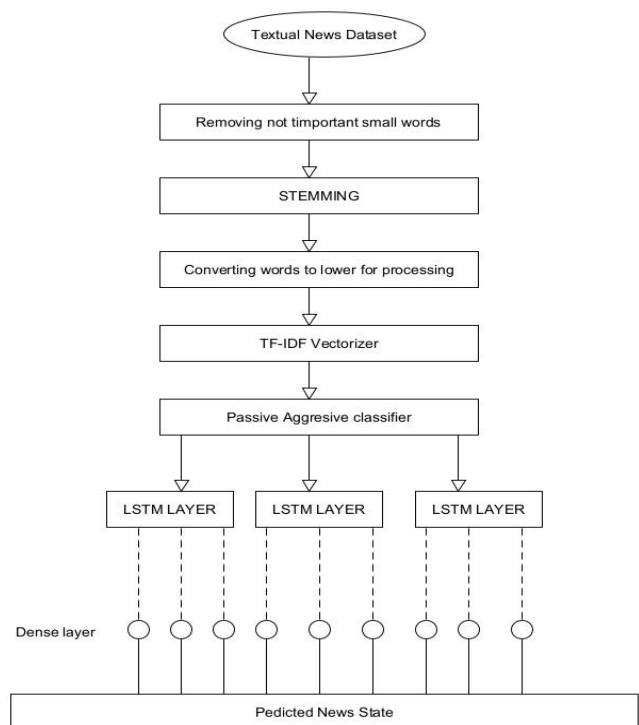
step 1: Textual dataset is prepared to identify fake news from it.

step 2: Using Stopwords library small words are removed from the   dataset. words like he, she, is etc.

step 3: in this step word from sentence are reduced to a root by removing inflection through dropping unnecessary characters, usually suffix.

step 4: each word is now converted to lower case for further analyzing.

step 5: TF-IDF vectorizer is used to extract features.

step 6: PA algorithm takes features for predication.

step 7: LSTM layer takes PA output as input in units and provide node to next layer.

Step 8: Dense layer is the last layer which provide final output of the probability of the each news of dataset.

Proposed solution consist of important steps 2 to 6 as preprocessing. it uses TF-IDF for tokenizing a collection of text document along with building vocabulary of pre-existing words. the main reason behind text preprocessing is to consider each document as a feature vector. PA algorithm takes that feature as input to categorize into passive and aggressive. this data feed into LSTM layer passed to a fully dense layer to produce single stance as final output.

VIII. Conclusion

with the inherent nature of internet, fake news gets circulated effectively without being identified as fake. Detection of fake news is not an easy task because it it always changing we cannot write the static code for capturing such fake news. so in order to enhance a fake news detection mechanism we should use machine learning techniques for dynamix solutions.

IX. Tools and Technologies

[1]   Multi-Source Multi-Class Fake News Detection Hamid Karimi, Proteek Chandan Roy, Sari Saba-Sadiya, and Jiliang Tang https://www.aclweb.org/anthology/C18-1131.pdf

[2]   Fake News Detection on Social Media: A Data Mining Perspective. kai shu, amy sliva, suhang wang, jiliang tang, huan Liu https://dl.acm.org/doi/abs/10.1145/3137597.3137600

[3]   Fake News Detection Using Sentiment Analysis Bhavika Bhutani Neha Rastogi Priyanshu Sehgal Archana Purwar **DOI:** 10.1109/IC3.2019.8844880

[4]   Unsupervised Fake News Detection on Social Media: A Generative Approach Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan   Wu,   Huan   Liu,   DOI: https://doi.org/10.1609/aaai.v33i01.33015644

[5]   The Role of User Profiles for Fake News Detection: kai shu, xinyi shou, suhang wang, reza zafarani, huan liu. https://dl.acm.org/doi/abs/10.1145/3341161.3342927

[6]   Supervised Learning for Fake News Detection, Julio C. S. Reis, Andre Correia, Fabrıcio Murai, Adriano Veloso, and Fabrıcio Benevenuto, **DOI:** 10.1109/MIS.2019.2899143

[7]   Fake News Pattern Recognition using Linguistic Analysis, Amitabha Dey; Rafsan Zani Rafi; Shahriar Hasan Parash; Sauvik Kundu Arko; Amitabha Chakrabarty