Malayalam Fake News Detection using Machine Learning

Anit Sara Santhosh
Department of Computer Application
Amal Jyothi College of Engineering Kanjirappally, India
anitsarasanthosh@mca.ajce.in

Dr. Bijimol T K
Department of Computer Application
Amal Jyothi College of Engineering Kanjirappally, India
tkbijimol@amaljyothi.ac.in

***Abstract* — Today everyone can access internet because of that there are different sources for online news. The news from social media is playing a huge role in controlling our life. This news can be fake or real. As common man to some extent users are unable to find out whether news is real or fake. Here users are trying finding a solution for this problem. The NLP techniques are used to detect the fake news from different sources. To build the model the dataset and machine learning algorithms are used. To detect accuracy of the model machine learning algorithm is used.**

Keywords—fake news, real news machine learning, count vectorizer, TF-IDF vectorizer, passive aggressive algorithm

I. INTRODUCTION

User gets news from different sources. This news can be fake or real. The use of fake news can manipulate our thinking skills and feeling. These types of fake news are published to increase the number of readers. The headlines of this type of news will make the readers more curious about it and by increasing the curiosity they will make people read this news, they will get more revenue. The fake news contents are increasing as the day passes, it is only because of the reason that users are depending a lot on social media sites. Users are also able to download and share these types of news articles; this will increase the complications in identifying the news whether it is fake or real. The people who are producing fake news are influencing our behavior and ability to make decision. This can affect us in the time of Artificial intelligence because there is a feedback loop. That is, one's user started reading these type of fake news the system will automatically show us the fake news only. This will produce a high impact on us and our society where user lives. This system checks that the Malayalam news given as input is fake or real. This fake news is always misleading us. By making these type of fake news the agencies can earn revenue. In this paper, users are developing a system that will accurately identify whether a new is fake or real. User should need a correct idea of what fake news is. To detect user, need to use different machine learning and natural language processing techniques.

The method for detecting fake news that uses:

- Text Preprocessing: Users are analyzing the test and removing the stop words.
- Encoding of the text: using the TF-IDF. This turns the raw document into matrix. Then fit and transform the vectorizer based on the train set.
- PassiveAggressiveClassifier: This is an online learning algorithm. This algorithm will passive for correct classification and aggressive for wrong classification. Users are also testing the accuracy of our model.

II. LITERATURE REVIEW

Niall J Conroy , Victoria L Rubin , Yimin Chen[1] propose classification of many ways for truth assessment rising from 2 main classes. they're linguistic cue approaches with machine learning and network analysis approaches for sleuthing faux news.

Mykhailo Granik, Volodymyr Mesyuran[2] develop a straightforward approach for faux news detection employing a naive theorem classifier. This approach is tested on a collection of information from Facebook news posts. They declare Associate in Nursing accuracy of seventy four. The speed of this model is sweet however not the simplest as a result of several different models are best. The models used completely different other classifiers.

Hadeer Ahmed , Issa Traore, Sherif Saad[3] developed a faux news detection model that uses n-gram analysis and machine learning techniques. They developed by scrutiny completely different feature extraction techniques and classification techniques. the simplest performances are gained by victimization the options extraction technique (TF-IDF). They conjointly use Linear Support Vector Machine classifier that has Associate in Nursing accuracy of ninety two.

DSKR Vivek Singh , Rupanjal Dasgupta[4] developed a brand new public dataset of reports articles. They planned a text-processing based mostly machine learning approach for automatic identification of pretend news. This has Associate in Nursing accuracy of eighty-seven accuracies. It's found that this model is specializing in the showing feelings of the text.

William Yang Wang developed a brand new dataset for automatic faux news detection[5]. This corpus also can be used for outlook classification, argument mining, topic modeling, rumor detection and political human language technology analysis. The works during this space have used this benchmark. it's celebrated that this is often restricted to political info however others have integrated info from varied fields.

III. MOTIVATION

The increasing amount of people dependence on online news should be taken as a concern. This will make us difficult to understand which is fake and which real news is. This should not be happened. This consumption of fake news by human beings should be reduced alarmingly. This system

DOI: 10.5281/zenodo.6162145
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

will help in this. The fake news articles can be identified and this will help us to differentiate between fake and real news. This model will help the common man to easily detect the news by analyzing the text in the input given by them. In this system users are using a dataset to build a model. The model is trained using this dataset build using passive aggressive classifier. Then the model is used to classify the news.

IV. METHODOLOGY

The purpose of this paper is to find a Machine Learning model to identify whether news is fake or real using the given dataset with accuracy. The model should be able to classify correctly as Fake and Real News. The dataset is divided into two training and testing data sets. The model should be trained with more data to get accuracy in learning. The result from this model can be used to create a news classification. The TF-IDF Vectorizer, Passive Aggressive Classifier will be used to learn to detect fake news. The structure is:

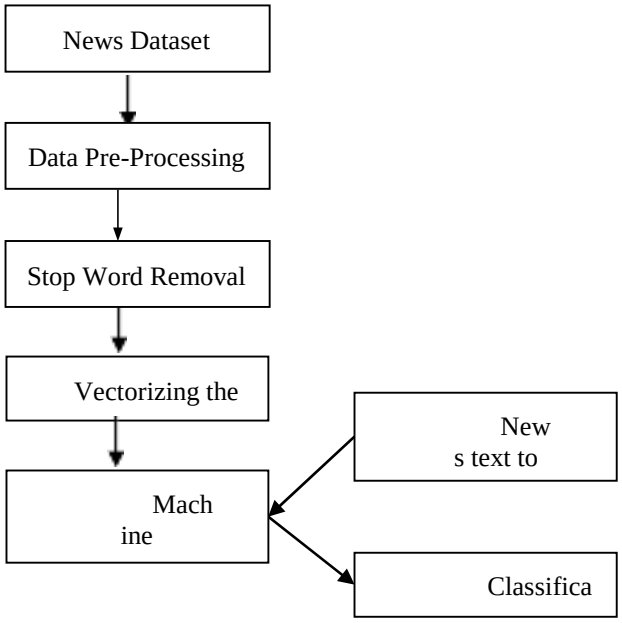

Figure 1: Malayalam Fake News Detection Methodology

The steps that require to be followed are:

1. Data Collection
2. Data Pre-processing
3. Model Building
4. Analyzing
5. Result

A. **Data Collection:** User have a dataset for malayalam fake news detection, the attributes of this dataset is used to train the model. There 316 rows and 2 columns in dataset. The columns are headings and label. The heading denotes the Malayalam news headings that can be fake or real. The label contains the labs that are FAKE and REAL. This is to differentiate the news between fake and real.

B. **Data Pre-processing:** This is the step that is done in every machine learning method. This process includes data cleaning, data transformation and data reduction. This all are done to make the data more effective. The data can be processed to make our model more accurate. So that the classification will be correct.

1. **Cleaning Data –** The data is cleaned at first. That is here users are removing the unwanted entries. The noisy data can also be removed from the dataset.
2. **Cleaning Data –** The data is cleaned at first. That is here users are removing the unwanted entries. The noisy data can also be removed from the dataset.
3. **Removing Stop words –** The stop words are removed from the datasets. These are the most frequently appearing that don't have much importance in the text. To achieve more accuracy to the model this can be done.
4. **Splitting of Data –** After cleaning the data the data is then split into training and testing datasets. After this the data chosen for training is used to train our model.

C. **Machine Learning:** This is used to detect the fake news. The machine learning algorithms are used to detect fake news, that will check the news given. There are different classifiers present to detect the fake news. The classifier accuracy relies on how it is trained. A classifier should need high accuracy. The steps are:

1. **CountVectorizer-** It can be used to transform text into vector based on the count of each word in the text. This will represent the words in a matrix format. In each cell there will the number of occurrence of each word in the text.
2. **Tf-Idf Vectorizer –** This is the commonly used method to measure the relevance of a word in a text. That is this can be used to convert a text into representation of numbers. This is done to derive features of text depending on its occurrence. The word with greater value has more importance. The equation is:

$$tf(w) = doc.count(w) / total\ words\ within\ the\ doc$$

In another context there will be words with high frequency but this will have less importance in the text string. This effect should be reduced for that user use inverse document frequency that is Idf. The equation is:

$$idf(w) = log(total\ number\ of\ documents) / (number\ of\ documents\ contacting\ word\ w)$$

Tf-Idf can be find out using,

$$tf\text{-}idf(w) = tf(w) * idf(w)$$

3. **Passive Aggressive Classifiers –** This algorithm is a type of machine learning algorithm. It is used for learning a large amount of data. It belongs to online learning algorithm. In the online learning algorithm it will take the training example and use that to

DOI: 10.5281/zenodo.6162145
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

update classifier and the example is no more used. This passive-aggressive algorithm does not need learning rate. They look similar to perceptron model. This algorithm is passive if the identification goes correct and the model will be taken. There won't be to any extent further changes to the model. This formula is aggressive if the identification goes wrong. there'll be some changes created to the model. There will be some changes made to the model to make it right.

4. There are some parameters used this include a regularization parameter, maximum number of iterations (max_iter) and a stopping criterion (tol).

V. BUILD MODEL

The model building is the main step in the fake news detection. While building the model user use the algorithms. The steps involved are:

1. Import the packages that are necessary.

```
In [2]: import os
        os.chdir("C:/Users/anit.sara/Jupyter/Fake_News_Detection-master")

In [3]: import pandas as pd
```

```
In [8]: from sklearn.model_selection import train_test_split
        from sklearn.feature_extraction.text import TfidfVectorizer
        from sklearn.linear_model import PassiveAggressiveClassifier
        from sklearn.metrics import accuracy_score, confusion_matrix
        from flask import Flask, render_template, request
        import pickle
        import pandas as pd
        from sklearn.model_selection import train_test_split
```

2. Add the data into a DataFrame, then get the shape of data.

```
In [4]: dataframe = pd.read_csv('train1.csv')
        dataframe.head()
```

Out[4]:

	headings	label
0	(ഫോട്ടോയിൽ ഇടതുവശത്ത് സണ്ണി ലിയോണാണോ	FAKE
1	തമിഴ്നാട്ടിൽ അർധരാത്രി മുതൽ പെട്രോൾ വില 65 രൂപ	FAKE
2	സൗരയൂഥത്തിന് അടിയിൽ ഭയാനക ശബ്ദം കേട്ടെന്ന വാ	FAKE
3	ലൈവിനിടെ പൊട്ടിച്ചിരിച്ച് ഉദ്യോഗസ്ഥരെ തിരിച്ചുവിളിച്ച	REAL
4	ഇറാനുമായി കനത്ത പോരിന് ശ	REAL

```
In [5]: x = dataframe['headings']
        y = dataframe['label']
```

3. Then split the dataset into training and testing datasets.

```
In [9]: x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.1,random_state=0)
        y_train
```

4. Initialize TfIdfVectorizer with stop words and set max_df as 0.7. That is the terms with a frequency higher than this max_df will be removed.

```
tfvect = TfidfVectorizer(stop_words='english',max_df=0.7)
tfid_x_train = tfvect.fit_transform(x_train)
tfid_x_test = tfvect.transform(x_test)
```

5. Then, fit and transform train and test set.
6. Initialize PassiveAggressiveClassifier.

```
classifier = PassiveAggressiveClassifier(max_iter=50)
classifier.fit(tfid_x_train,y_train)
```

7. Then identify on test set and calculate the accuracy of the model.

```
y_pred = classifier.predict(tfid_x_test)
score = accuracy_score(y_test,y_pred)
print(f'Accuracy: {round(score*100,2)}%')
```

The accuracy is 98.4%.

8. Print the confusion matrix.

VI. RESULT

The result shows that the Malayalam news can be classified as fake and real news based on the dataset given. The Passive Aggressive provides correct result with high accuracy. The dataset contain news from different areas that are real and fake. This is classified and the fake news was correctly detected by the model.

DOI: 10.5281/zenodo.6162145
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

VII. CONCLUSION

This paper provides a method for detecting Malayalam fake news using PassiveAggressiveClassifier. This is using best features to detect the news that are fake. Here user implemented a solution by first processing the data, then using TF-IDF to derive some features of the dataset. Then user has used the PassiveAggressiveClassifier to build a model to classify the fake news.

The work has a lot of future scope. The dataset can be extended and can use any other online learning algorithms or tools to detect the Malayalam fake news. Also, real time Malayalam fake news detection in videos can be done in future. The other scope is that key sources of the fake news can be found out using graph theory and machine learning techniques.

REFERENCES

[1] Niall, Victoria, and Yimin. Automatic deception detection: ways for locating fake news. Proceedings of the Association for information processing and Technology, 52:1–4, 2015.

[2] Mykhailo Granik and Volodymyr Mesyura. retend news detection mistreatment naive bayes classifier. In 2017 IEEE initial Ukrayina Conference on Electrical and Computer Engineering (UKRCON),. IEEE, 2017.

[3] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detection of on-line pretend news mistreatment n-gram analysis and machine learning techniques. Springer, 2017.

[4] DSKR Vivek Singh and Rupanjal Dasgupta. automatic faux news detection pattern linguistic analysis and machine learning.

[5] William rule Wang. " liar, slicker pants on fire": a replacement dataset for pretend news detection. arXiv preprint arXiv:1705.00648, 2017.William Yang Wang. " liar, liar pants on fire": a replacement dataset for fake news detection. arXiv preprint arXiv:1705.00648, 2017

DOI: 10.5281/zenodo.6162145
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam