International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 10, Issue 7 [July. 2021] PP: 12-15

Fake News Detection Using Machine Learning

A Siva Krishna Reddy[1], Y Lavanya[2], K Naveena[3], Y Naga Nikhil[4], R Dileep Kumar[5], K N Brahmaji Rao[6]

[1]*Associate Professor, Department of Computer Science and Engineering, Raghu Institute of Technology, Vizag, India*
[2]*Student, Department of Computer Science and Engineering, Raghu Institute of Technology, Vizag, India*
[3]*Student, Department of Computer Science and Engineering, Raghu Institute of Technology, Vizag, India*
[4]*Student, Department of Computer Science and Engineering, Raghu Institute of Technology, Vizag, India*
[5]*Student, Department of Computer Science and Engineering, Raghu Institute of Technology, Vizag, India*
[6]*Associate Professor, Department of Computer Science and Engineering, Raghu Institute of Technology, Vizag, India*

ABSTRACT*: In recent years, due to the low cost of internet and the wide usage of social media platforms, fake news is spreading quickly on social media. On one side news can be found easily on social media with low cost and easy access but on the other side with the fast spread of fake news there is a negative impact on society. Fake news can impact a person's intuition. So, it is important to detect if a news is fake or real timely to reduce the negative impacts on society. We use Machine Learning Natural Language Processing techniques to build a classifier to detect fake news using python. For better performance we are using feature selection method integrated with k-means clustering algorithm. Feature selection is integrated with k-means clustering in-order to reduce the dimensionality of the dataset. Since the dimensions of the dataset are reduced, computational complexity will be reduced and in-turn improves the performance Support vector machine classifier is used to label the data as fake or Real.*

Keywords: *Classification, Fake News, Feature Selection, K-means Clustering, Support Vector Machine.*

Date of Submıssıon: 18-07-2021 Date of Acceptance: 03-08-2021

I. INTRODUCTION

Fake news is nothing but a kind of false information or misinformation or rumour which spreads on social media platforms. A fake news can be spread intentionally by an individual user or a group of users holding an account in social media. Fake news can be of any form. A fake news can be an article or an image or a video. A fake news may impact and mislead a person's decision-making ability. As per a survey conducted by Social Media Matters and Institute for Governance, Policies and Politics, every 1 in 2 Indians receives fake news through WhatsApp and Facebook. So, researches on fake news detection have a lot of attention in recent times. The existing system can label the data as fake or real using feature extraction and a data classifier but the vast usage of social media platforms results in huge amount of data. As the data is huge, the computational complexity is increasing. In-order to reduce the data dimensions, redundant features need to be eliminated and also the irrelevant features need to be removed. Here we use an NLP technique TF-IDF vectorizer to extract features and the features are clustered using k-means clustering algorithm, then feature selection technique is applied on the clusters of features to get the most optimal features to reduce computational complexity. Finally, an SVM classifier is used to build the model.

II. LITERATURE REVIEW

In this paper [1] machine learning and python are used for fake news detection. Natural language processing techniques are used to classify the data. This model is built on NLP techniques, count vectorizer and TF-IDF vectorizer. Naive Bayes and Support Vector Machine classifiers are used to classify the data based on features extracted using TF-IDF vectorizer.

In this paper [2] the application of Natural Language Processing and Machine Learning techniques are explored to identify fake news accurately. Data is pre-processed and feature extraction is applied on them. A fake news detection model is built using four techniques and compares the accuracy of techniques which are Naive Bayes, Support Vector Machine (SVM), neural network and long short-term memory (LSTM) to find the model best fits it.

In this paper [3] the existing system with feature extraction is developed using feature selection method integrated with k-means clustering algorithm to reduce the redundant and irrelevant features to improve accuracy and to reduce computational complexity.

In this paper [4], we present a comprehensive review of detecting fake news on social media, including fake news characterizations on psychology and social theories, existing algorithms from a data mining perspective, evaluation metrics and representative datasets.

III.ARCHITECTURE OF PROPOSED SYSTEM

This paper aims at studying the fake news detection using Machine Learning algorithms and Natural Language Processing techniques. It takes training data as input, trains the model and evaluates it using the testing data. We can see the entire overview of the model from the below figure.

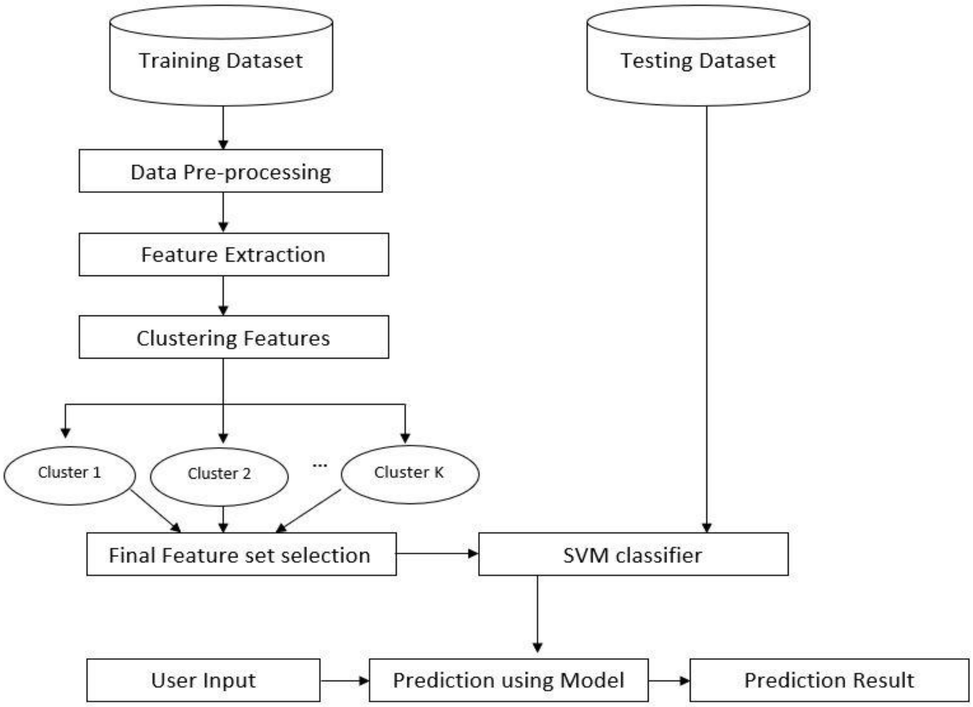

Fig 1 System Architecture

Fig 1 clearly illustrates the construction of the classifier model. Training dataset is pre-processed and then features are extracted from it. Then all the similar features are clustered into different clusters. Topmost features are selected from each cluster using feature selection techniques. Then these final set of features are trained to the SVM classifier to predict the user input as FAKE or REAL.

IV.METHODOLOGY

A. Pre-Processing

Data pre-processing involves cleaning the data and removal of noisy data. Removing noisy data helps in improving performance of the classifier. Stop words like ‘a’, ’an’, ’the’ etc. are removed as they are not significant to process the text. NLP techniques are used to pre-process the data. The entire text or corpus is converted into lower case and special characters are also removed. It also involves expanding abbreviation, stemming which converts each word into its root word.

B. Feature Extraction

Features are extracted from the corpus and the TF-IDF Vectorizer is used to find the significance of each feature. TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction. Term frequency refers to the number of times a given term will appear in a document and Inverse document frequency measures the weight of each word in the document, i.e., if the word is common or rare in the entire document. Python based libraries are used for extracting features using TF-IDF Vectorizer.

C. Clustering Features

The features extracted are clustered based on similarity. Here we use the K-means clustering algorithm to cluster the similar features. The value of K is selected in such a way that it best fits the model. The similarities among the features can be calculated using Euclidean distance, Pearson's correlation coefficient etc. Then the clusters are formed using k-means clustering. K centroids are initialized randomly and then other data entities join the nearest cluster centroids to form new clusters with new centroids. This process continues until each data entity (feature) is allocated to its closest cluster centroids. In each iteration, the centroids of clusters are updated with their new entities and this continues until no more improvement happens.

D. Feature Selection

Within each cluster, the best features are selected based on the TF-IDF score which is calculated during the Feature Extraction. So, the dimensions can be reduced as the redundant features and the irrelevant features are not considered for classification. The dimensions of data which is used to train the model is reduced to k×m size, where k is the number of clusters and m is the number of features. Here we use the feature selection technique univariate feature selector Select Percentile with f_classif as score function. f_classif is an ANOVA F-Value based label or feature classification task.

E. Train SVM Classifier

After creating the final feature set, fake news can be detected using a classifier. SVM Classifier is a supervised learning algorithm used for classification. An SVM classifier works by being trained with data already classified into two different classes. SVM is used to separate fake news data with hyperplanes and extend it to non-linear boundaries. The model is built after training with a training dataset. Finally, the model is used to predict the user input as real or fake.

V. RESULTS

After training the model with 80% of the dataset, the model is tested using the remaining 20% dataset to find the accuracy of the model. The accuracy of the existing system i.e., feature extraction based with SVM classifier is about 83%. But the accuracy obtained using feature selection with clustering is about 93%. But the model works well only with the data related to the training dataset.

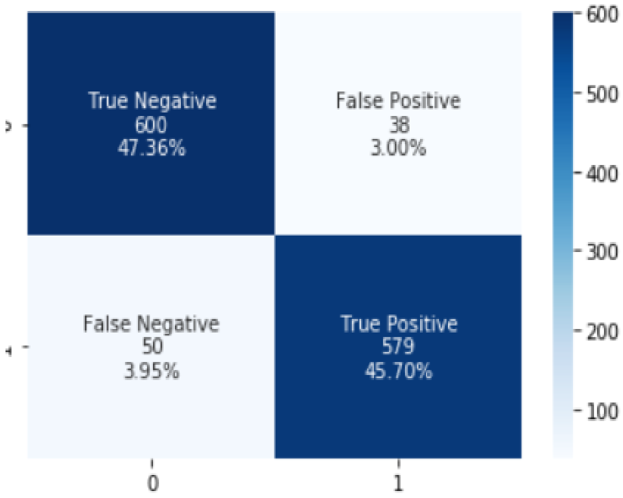

Fig 2 Confusion Matrix for SVM Using Feature Selection

Table 1
Comparison of Results

Algorithm Name	**Accuracy**	**Technique**
Support Vector Machine	93.05%	Feature Selection
Support Vector Machine	82.3%	Feature Extraction
K Nearest Neighbors	79.2%	Feature Extraction
Decision Tree Classification	82.7%	Feature Extraction

From the Table 1 it is clear that feature selection technique has given best accuracy when compared to feature extraction technique. And also, SVM classifier has given best accuracy when compared to other algorithms using feature selection approach.

VI. CONCLUSION

With the vast usage of social media, a huge number of people are consuming news from it. Social media is the main source for spreading fake news with a fake identity to mislead users. Social media platforms like Facebook, Twitter etc. almost allow anyone to share their thoughts and stories to the world. Most of the people before checking the source of content that they view online tend to share it, which results in spreading of fake news quickly or even going viral. Using fake news detection using appropriate algorithms we can avoid all the misleading news, rumours and hoaxes. A fake news detector plays a vital role in preventing the spread of fake news. The classifier we build helps to predict the news as fake or real.

ACKNOWLEDGEMENT

F A Authors like to express our sincere gratitude to our project guide [1]Dr. A. Siva Krishna Reddy M.E., Ph.D.**,** Associate Professor and coordinator [6]Dr. K N Brahmaji Rao M.Tech., Ph.D., Associate Professor, Department of Computer Science & Engineering, Raghu Institute of Technology for his whole hearted guidance, encouragement and continuous support throughout this work.

REFERENCES

[1]. Shruthi S. Shetty, K. B. Shreejith, Deekshitha, Dhanusha, K. B. Gagana, Fake News Detection Using *Naive Bayes and Support Vector Machine Algorithm*, International Journal of Research in Engineering, Science and Management Volume-3, Issue-6, June-2020

[2]. Kasra Majbouri Yazdi, Adel Majbouri Yazdi, Saeid Khodayi, Jingyu Hou, Wanlei Zhou, Saeed Saedy, Improving Fake News Detection Using *K-means and Support Vector Machine Approaches*, World Academy of Science, Engineering and Technology International Journal of Electronics and Communication Engineering Vol:14, No:2, 2020

[3]. Poonam Tijare, A Study on Fake News Detection Using *Naïve Bayes, SVM, Neural Networks and LSTM***,** Journal of Adv Research in Dynamical & Control Systems, Vol. 11, 06-Special Issue, 2019.

[4]. Kai Shuy, Amy Slivaz, Suhang Wangy, Jiliang Tang, and Huan Liuy, Fake News Detection on Social Media:*A Data Mining Perspective*