Newfangled Approach for Fake Content Detection

Shivani Santoki[1], Nivid Limbasiya[2]

[1]Student, Department of Computer Engineering, V.V.P. Engineering College, Rajkot, Gujarat, India
[2]Professor, Department of Computer Engineering, V.V.P. Engineering College, Rajkot, Gujarat, India

Abstract - *These days' internet based life is generally utilized as the source of data due to its simple to get to. Even so, it also has negative side because of the wide extension of fake news, i.e. news with misleading information. The issue has been drawn nearer in this paper from Natural Language Processing and Machine Learning points of view. To discover whether the news is fake and genuine is a difficult task. The truth of any statement often cannot be assessed by computers alone, so efforts depend on collaboration between humans and technology characterizes hundreds of popular fake and real news measured by shares, likes reactions, and comments, emoticons on Social Media from two ways: headlines and content.*

Various algorithms are proposed to detect fake news based on different features, domains and type of news. We propose a Gradient Boosting classifier that captures real time dataset process it and which identifies sentiment efficiently. This approach improves performance of system by increasing accuracy, precision, F-score Performances of different machine learning algorithms in terms of accuracies and F1 scores are compared. The motivation behind the exploring work is to come up with a solution that can be utilized by individuals to detect and filter out sites containing false and misleading information. We compared different algorithms with our proposed system and it gives best result out of them.

Key Words: **Fake News Detection, Efficient XGB, Natural Language Processing, Machine Learning**

1. INTRODUCTION

Currently, most of the people are using social media like twitter, Facebook and micro blogging sites. They share their opinions, feelings for particular scenario through comments, review. Volume of data generated daily is very large. So it is necessary to analyses the data for gaining information from that. Fake News Detection is used for differentiate fake news as well as real news. It refers to use natural language processing, text analysis, extract and study affective information.

Fake news detection is a foremost and technically difficult problem. Endeavour to tackle the expand information, several fact-checking websites. While speaking, many different clues like gesture, hand movement, rolling eyes, and facial expressions are available. But in textual data this type of clues are not available and more of informal nature with vocabulary of slang words and abbreviation. There is lot of data are generated in daily bases. These huge amounts of data can be processed through boosting, which is highly effective and widely used machine learning method. The evaluation is carried out for these standard datasets with set of feature extracted from the headlines and contents.

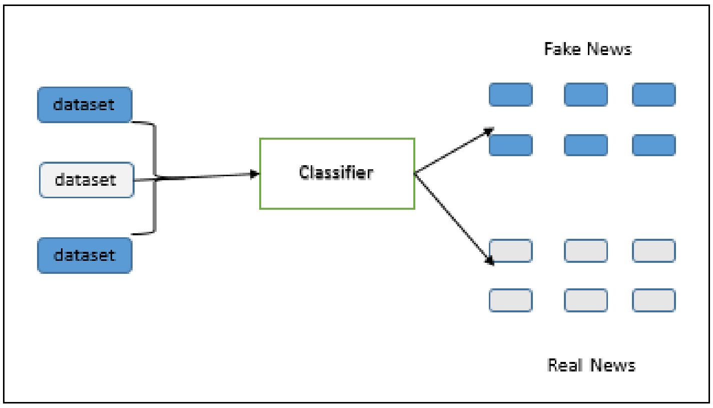

Fig 1. Fake News Detection

There are different models that are proposed, it is observed that many pre trained models are used for fake news detection. We proposed a model name as efficient XGboost.

The contribution of this paper is as follows:

1) We have transfer data into classifier for classify dataset in fake news and real news classes. There are different procedures done on data for result. It uses Machine Learning model and uses features like 1) n-grams count feature uses for counting occurrences in title and body of news.

2) Word embedding uses for vector representation of words, which replaces each word with real value vector.

3) So different preprocessing like convert raw data in structured data, handling Null values, feature selections is done.

In section 2. Related work, section, 3. Describe the Methodology and architecture of Proposed Model. Section 4. Experimental result, Section 5. Conclusion, Section 6. Acknowledgement.

2. RELATED WORK

Many Approaches have been used earlier to solve the Fake News problem. Different Machine learning models are discussed here in the section. All these have still some limitations like sentiment analysis can also explore, explore word2vector for word embedding, expand feature only for textual data but also can do for image as well as video. And also these methods are incompatible with different large datasets. So we have developed methods name as Efficient XGBoost for best result of fake news detection. Our main aim is to provide better result of detection with high accuracy. The Efficient XGBoost classifier is used here for classify the results.

Monther Aldwairi and Ali Alwahedi Introduced solution that can be utilized by users to detect and filter out sites which have false as well as misleading information. Select features from title and post to find fake posts. They had use WEKA machine learning for validate data and use different classifiers for evaluation and get high accuracy result with logistic classifier [1]. Kai Shu. Introduced a system FakeNewstracker for understanding and detection of fake news and it will also collect data. They used visualization techniques. [2] Kuai Xu presented that basically put in term frequency - inverse document frequency (tf-idf) and Latent Dirichlet allocation (LDA) subject modeling is inefficient in detecting fake news, so exploring document similarity with the term and word vectors is a better way for predicting fake and real news [3]. Elshrif Elmurngi, Abdelouahed Gherbi Introduced movie reviews into different groups to identify fake and real data and they used sentiment analysis concept. They utilized different text classification methods and supervised machine learning algorithm [4] Oluwaseun Ajao Presented their hypothesis by comparing with the state-of-the-art baseline text-only fake news detection methods that do not study sentiments. They did experiment on standard Twitter fake news dataset. They compute the classification of the labeled dataset using a series of machine learning algorithms [5]. Bhavika Bhutani Presented a new solution for fake news detection which incorporates sentiment as an important feature to raise the accuracy. [6]. Ammu Kuriakose Introduced system is named ALIKAH fake news detection system. They used neural network as the classifier. [7]. Basant Agarwal proposed system which automatically classifying sentiment of Twitter and give result in positive and negative. They also used hybrid structure on standard datasets. [13]

3. METHODOLOGY

The Proposed system have the following steps: 1) pre Processing 2) Feature Extraction 3) Efficient XGBoost for fake news and real news.

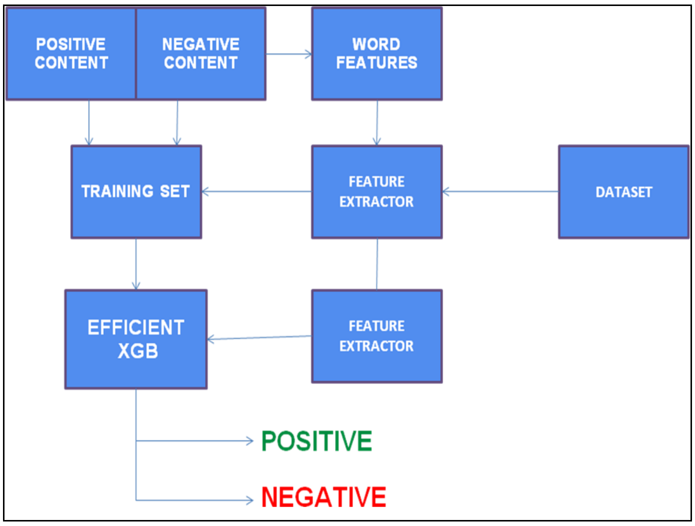

Fig.2. Architecture of proposed system

Fig 2. shows the architecture of our proposed system for best classification of fake news and real news. In the first step the data is converted into structured dataset from the raw dataset. Then for the best result of classification there are different steps are followed that are get data, data preprocessing, feature extraction and classifier.

3.1 Data Preprocessing

In real word data have no proper format, it contains noises, may be missing values and also unusable format which cannot be directly used by models. So main task of data preprocessing is to clean the data and make it suitable for model to get better accuracy. There are different steps are available for data preprocessing like

Get the Dataset

Import libraries of python like Numpy, Pandas, Matlotlib etc.

Import Data: import dataset in .csv file because it executes fast

Finding Missing Values: here we can either delete that particular row if it has NULL value when enough sample of data available then only use this method otherwise we loss the important data or if we have numeric data then we can take mean, median, mode of feature and replace it with missing values.

Splitting Dataset in to Training and Test sets: Generally, we have to define data in some ratio which depend on shape and size of dataset.

3.1 Feature Extraction

We have used feature extraction for reduce number of resources for processing without losing important information where data is become in more manageable groups for further processes. There are different feature extractions are available as below.

3.2.1 Number of Words

It is used for count the number of STOP Words because generally we remove it in NLP Problem but it might be give us some extra information which we might be losing it.

3.2.2 Number of Uppercase Words

Anger or rage is quite often expressed by writing in UPPERCASE words which makes this a required operation to spot those words.

3.2.3 N-grams

N-grams are the combination of multiple words used together.

3.2.5) Term Frequency: Term frequency is simply the ratio of the count of a word present in a sentence, to the length of the sentence.

3.2.4 Number of STOP Words

It is used for count the number of STOP Words because generally we remove it in NLP Problem but it might be give us some extra information which we might be losing it.

3.2.5 Term Frequency

Term frequency is simply the ratio of the count of a word present in a sentence, to the length of the sentence.

TF = (Number of times term T available in the row) / (number of terms in that row)

3.2.6 Inverse Document Frequency

We have used it for word not much use if it's available in all the documents.

IDF = log(N/n), where, N is the total number of rows and n is the number of rows in which the word was present.

3.2.7 Bag of Words

Bag of Words (BoW) refers to the representation of text which describes the presence of words within the text data. The intuition behind this is that two similar text fields will contain similar kind of words, and will therefore have a similar bag of words. Further, that from the text alone we can learn something about the meaning of the document.

4. EXPERIMENTAL RESULT & ANALYSIS

We have used four datasets for our experiment. Open source dataset having many articles from categories fake and reliable were selected [8], Kaggle dataset have 28000 plus articles [9], fake_real_news_dataset having two sections like headline and text of news [10]. For proposed model features are selected and apply on these datasets. Also dataset preprocessing and cleaning is one of the important task in natural language processing. The Performance matrix is used to calculate Precision, Recall, F1 and Accuracy. All of these are determined dependent on the true positive, true negative, false positive and false negative. The equations are as follows.

$$\mathrm{Pr}ecision = \frac{TruePositive}{TruePositive + FalsePositive}$$

$$\mathrm{Re}call = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F1 = \frac{2 * \mathrm{Pr}ecision * \mathrm{Re}call}{\mathrm{Pr}ecision + \mathrm{Re}call}$$

Table -1 Comparison of different model's accuracy on Open sources, Kaggle and George McIntire dataset.

Classifier	Open Source	Kaggle Dataset	George McIntire dataset
Efficient XGBoost	89.3	92.9	89.5
RF	81.2	86.63	82.6
SVC	62.9	63.55	62
KNN	62.5	72.54	73.2

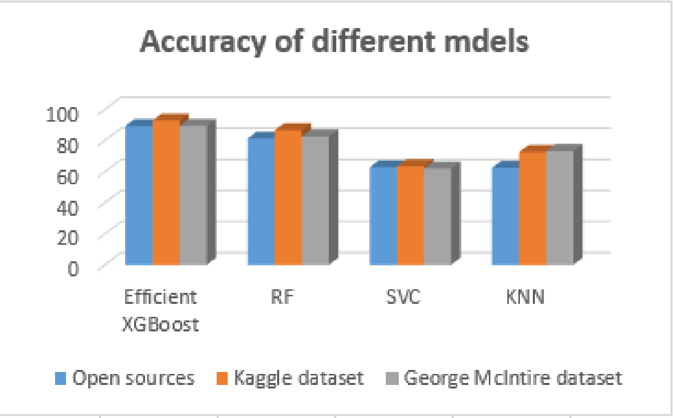

Table -1 Comparison of different word embedding model on different dataset.

Dataset	Precision	Recall	F1-Score
Open Sources	0.93	0,91	0.92
Kaggle Dataset	0.94	0.93	0.94
George McIntire	0.90	0.88	0.89

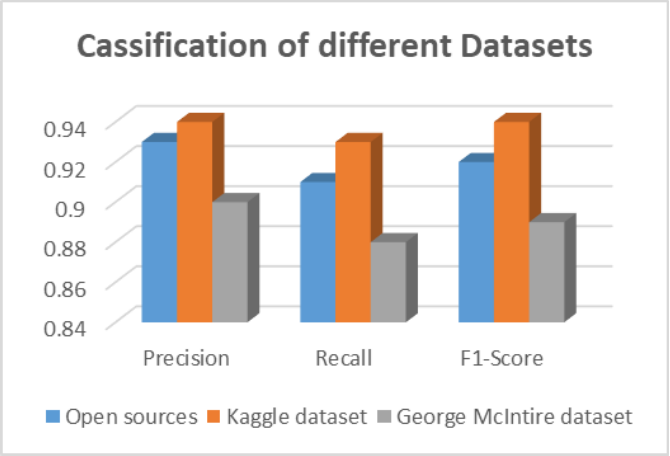

Fig.1 shows the comparison of different techniques which have been proposed earlier and comparison with our proposed method shows that proposed method which is Efficient Xgboost having highest results among state-of-the-art methods.

5. CONCLUSIONS

In this Research paper, we presented Efficient Xgboost algorithm for find fake content. We have compared with all the other existing methods, this method is only based on neural architecture and Machine learning approach. We have also test one more than one datasets. The experimental result shows that the proposed method achieves with very good performance. We have different preprocessing method like remove duplicates, generate ngram models for proper dataset from raw data. We have used different performance matrices name as accuracy, precision, F-score and recall for results. In this paper we have used different word embedding for similarity measurement like cosine similarity. We have used features like counter feature, readability, sentiment, wordtovector etc. Proposed model gives us better accuracy among state-of-the-art methods. Which is highest among previously used approach. The future work can be extended for more features generation for fake news detection.

ACKNOWLEDGEMENT

This research work was carried out with the help of faculty members of V.V.P. Engineering College. Special thanks to Professor Nivid Limbasiya for their guidance and support throughout the whole dissertation work. I also do not like to miss the chance to acknowledgment the endowment of all faculty members of department for the kind helping hands and collaboration during research work.

REFERENCES

1) Monther Aldwairi, Ali Alwahedi, " Detecting Fake News in Social Media NetworksThe 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks(EUSPN 2018)

2) Kai Shu, Deepak Mahudeswaran, Huan Liu1, "FakeNewsTracker: a tool for fake news collection, detection, and visualization" Springer Science+Business Media, LLC, part of Springer Nature 2018

3) Elshrif Elmurngi, Abdelouahed Gherbi(2017) , "Fake Reviews through Sentiment Analysis Using Machine Learning Techniques" In The Sixth International Conference on Data Analysis.

4) Shao, Y.: HCTI at SemEval-2017 task 1: use convolutional neural network to evaluate semantic textual similarity. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017), pp. 130–133 (2017)

5) Oluwaseun Ajao1, Deepayan Bhowmik2 and Shahrzad Zargari1,

“SENTIMENTAWAREFAKENEWSDETECTIONONON LINESOCIALNETWORKSFang” IEEE 2019

6) Bhavika Bhutani Neha Rastogi Priyanshu Sehgal Archana Purwar, “Fake News Detection Using Sentiment Analysis” IEEE 2019

7) Ammu Kuriakose, Dinnu Sebastian, Esther Mahima Mathew, Hannu Mathew, Er.Gokulnath G, “ ALIKAH- A Clickbait and Fake News Detection System using Natural Language Processing” IEEE 2019

8) Opensource Dataset. http://www.opensources.co/

9) Kaggle Dataset. https://www.kaggle.com/jruvika/fake-news-detection

10) GitHubRepository.https://github.com/GeorgeMcIntire/fake_real_news_dataset

11) XichenZhang⁎,AliA.Ghorbani , “Anoverviewofonlinefakenews:Characterization,detection,and discussion” 2019 Elsevier Ltd

12) Namita Mital, Basant Agarwal, Saurabh Agrawal, Pramod Gupta(2013), “A Hybrid Approach for Twitter Sentiment Analysis. In 10th International Conference on Natural Language Processing.