PUNE VIDYARTHI GRIHA'S COLLEGE OF ENGINEERING & TECHNOLOGY AND G K PATE (WANI )

INSTITUTE OF MANAGEMENT PUNE-09

Seminar Presentation on

# "An Ensemble Machine Learning approach through effective feature extraction to classify FAKE NEWS"

By

**Atharv Sunil Bobade**
**Roll no. : 9046**

Under the guidance of
"**Prof. S. G. Kamble**"


# Content

- Aim & Objectives of Project
- Introduction
- Literature Survey
- Proposed Methodology
- Data Cleaning
- Data Visualization
- Ensemble Model Selection
- Training of Model
- Experiment results
- Application
- Conclusion
- References

# Aim & Objectives of Project

## The Aim of this Seminar is

-> To learn How to find the features to train the Machine Learning Model
-> Using Machine Learning Technique to verify the Truthiness of Information

## Objective

-> Analyse the dataset.
-> Pre-processing of text.
-> Tokenization.
-> Extracting features from Text.
-> Train Machine Learning Model.
-> Evaluate performance of model.
-> Predicting the fake news.

# Introduction

- The concept of fake news has been in existence even before the emergence of Internet and other computational technologies.
- Dissemination of fake news and misleading information has always been used as a weapon to fulfil immoral objectives since ages.
- The advancement of technologies has enabled convenient access to authentic and falsified information even faster posing a real challenge.
- The spread of such fake news has extremely negative impact on target individuals and also the society at large.
- Fake information affects stock prices, choice of stock purchases, investment plans and even reactions to natural calamities and many.
- It is very important to stop circulation of misleading falsified information.

- The present study involves experimentation on two popular fake news datasets, ISOT and Liar datasets.
- Part of the proposed algorithm, 70% of dataset is used for training and remaining 30% is used to test the classification model using k-fold cross validation.
- The main contributions of this paper are :
  1. Using feature extraction to use the most significant features that influence the classification of fake news.
  2. Selection of an Ensemble model to achieve optimized accuracy in classification.
  3. Reduction of training time of the ensemble classifier.

# Literature Survey

1. **P.H.A. Faustini, T.F. Covões, Fake news detection in multiple platforms and languages, Expert Syst. Appl. (2020) 113503.**
   - Introduces fake news detection in multiple languages, namely Slavic, Latin, and German by selecting text features.
   - The experiments were carried out on five different datasets, namely TwitterBR, FakeBrCorpus, FakeNewsData1, Fake-OrRealNews, and btvlifestyle.
   - Finally, each dataset is fed to different classification algorithms such as KNN, Support Vector Machines (SVM), Random Forest (RF), and Gaussian Naive Bayes (NB). SVM and RF deliver better performance compared to conventional methods.
   - However, the proposed model does not achieve a higher accuracy rate and thus restricts its use in the identification of fake news.

2. **M.D. Vicario, W. Quattrociocchi, A. Scala, F. Zollo, Polarization and fake news: Early warning of potential misinformation targets.**
   - presents a new way of predicting fake news in social media well in advance.
   - the authors proposed new classifier with different features, such as semantic features, user-based features, structural features, sentiment-based features, and predicted features.
   - The research was carried on an Italian Facebook dataset with 300K official media news and 50 K posts from various blogs and websites providing false or incorrect facts.
   - The results show the proposed early detection of fake news achieved 77% training accuracy and 91% testing accuracy.
   - However, the fake news prediction can be enhanced by identifying the elements that negatively impact the information.

**3. Y. Liu, Y-F.B. Wu, FNED: A deep network for fake news early detection on social media**

- The status-sensitive crowd feedback acts as input during the process, and the Convolution Neural Network (CNN) is used as a classifier educated with positive and unlabelled samples
- Five-fold cross-validation is used for model validation.
- The experiment was carried out on two different datasets, namely Twitter and Weibo.
- The proposed model achieved a 90% accuracy rate within 5 minutes of the news spread.
- However, the authors have used a small dataset which limited the proposed work.

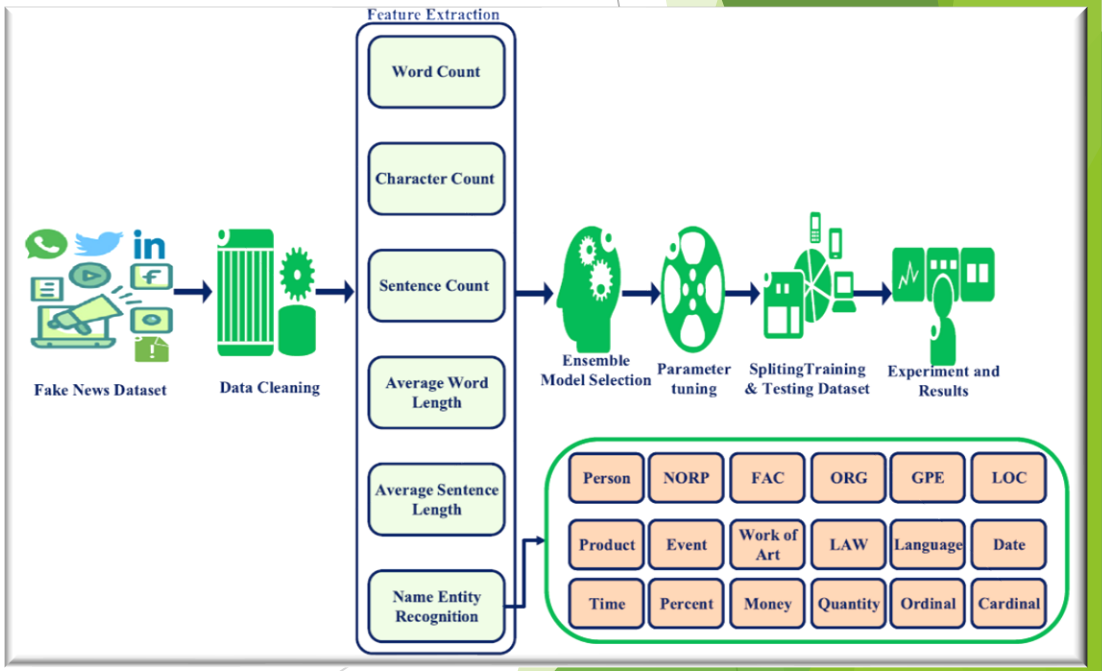
| Research Paper | Dataset | Contributions | Accuracy Attained | Challenges |
|---|---|---|---|---|
| P.H.A. Faustini, T.F. Covões, Fake news detection in multiple platforms and languages | TwitterBR, FakeBrCorpus, FakeNewsData1, FakeOrReal News, and btv lifestyle | Fake news detection in multiple languages | 79% | Does not attain higher accuracy |
| M.D. Vicario, W. Quattrociocchi, A. Scala, F. Zollo, Polarization and fake news: Early warning of potential misinformation targets | Italian Facebook dataset with 300K official media news, 50K incorrect information | Proposed a classifier with different features, such as semantic features, user-based features, structural features, sentiment-based features, and predicted features | 91% | Failed to show identifying the element that negatively impact the information |
| Y. Liu, Y.-F.B. Wu, FNED: A deep network for fake news early detection on social media | Twitter and Weibo datasets | CNN is trained with positive and unlabelled samples. Five-fold cross-validation is used for model validation. | 90% | Limited to a tiny dataset of 1,111 Twitter posts and 816 Weibo posts |

# Proposed Methodology

- In the proposed study, writers identified twenty-six (26) linguistic based textual features as listed in table.
- The identified features were extracted from the text. For detecting fake news, the state-of-the-art machine learning models were explored.
- Figure presents the architecture of proposed work. It consists of several phases such as data pre-processing, feature selection, ensemble-model-selection, hyperparameter tuning and training of the model.

Features extracted from text.

| Feature name | Data type | Feature name | Data type |
|---|---|---|---|
| Person | Numeric | NORP | Numeric |
| FAC | Numeric | Organization | Numeric |
| GPE | Numeric | Location | Numeric |
| Product | Numeric | Event | Numeric |
| Work of Art | Numeric | Law | Numeric |
| Language | Numeric | Date | Numeric |
| Time | Numeric | Percent | Numeric |
| Money | Numeric | Quantity | Numeric |
| Cardinal | Numeric | Ordinal | Numeric |
| word_count | Numeric | char_count | Numeric |
| sentence_count | Numeric | avg_word_length | Numeric |
| avg_sentence_length | Numeric | polarity | Numeric |
| avg_sentence_length | Numeric | sentiment_score | Numeric |



## The brief description of the steps involved are as follows:

- Popular fake news datasets i.e. liar and ISOT were identified and explored for the proposed approach.
- To remove the noise from datasets, necessary pre-processing was done.
- After text pre-processing, tokenization is performed to convert the larger text in to words or in small lines.
- The major part of this research is extracting the features from text and then use these features for fake news detection instead of text.
- The extracted features are then passed to the state-of-the-art machine learning algorithms like ensemble decision tree, random forest and extra tree classifier to train the model.
- Various evaluation metrics are used to evaluate the performers of our proposed model.

# 1.Data Cleaning

- For this study explored two popular datasets i.e. Liar and ISOT.
- Liar dataset consists of labelled short statements covering various news topics that have been labelled manually.
- ISOT dataset is created by University of Victoria. It contains 23,481 fake news articles and 21,417 true news articles respectively.

## Data Pre-processing

- In this phase, the given datasets were pre-processed to remove the noise such as stopwords, punctuation marks, html tags, url, emojis, etc.
- Pre-processing was done using NLTK toolkit which is an open-source and widely used NLP library. It comes with inbuilt functions and algorithms such as nltk.tokenize method (for tokenising text), nltk.stem.porter.PorterStemmer method.

The pre-processing of datasets is carried out as follows

- Tokenization.
- Stop Words Removal
- Stemming
- Features Extraction

- **Tokenization:** is the process of splitting the text/string into the list of tokens and is considered as the first step in natural language processing before feature extraction process. nltk.tokenize method (an inbuilt function in nltk library) is used in this work for tokenization.

- **Stop Words Removal**: After tokenizing the text, this step consists of removing the stop words. Stop words are insignificant words in a language that create noise when used as features in text classification. These words are frequently used in sentences to connect different words or to assist in the sentence structure. It consist of common words such as a, about, an, are, as, at, be, by, for, from, how, in, is, of, on, or and so on.

- **Stemming:** Stemming is a process of reducing the words to its root (also known as lemma). For example, the words such as running, ran, and runner will be reduced to its lemma which is a word run. For this purpose, The porter stemmer algorithm is used.

- **Features Extraction:** Twenty-six features were identified for this study. The reason for selecting the less number of features was due to the fact that irrelevant features decreases the accuracy of the models and increases the cost of training process.

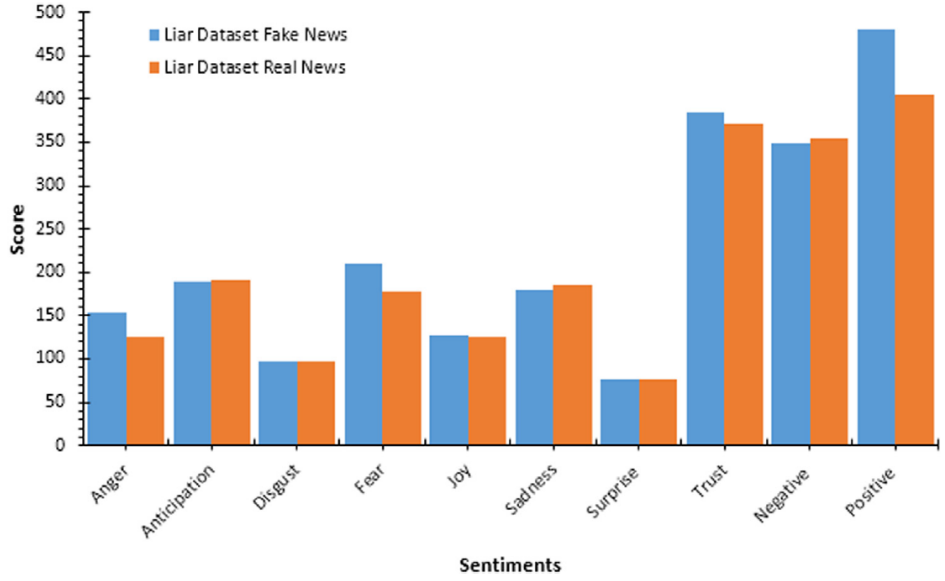



**Fig. 5.** Liar dataset fake and real news emotions.

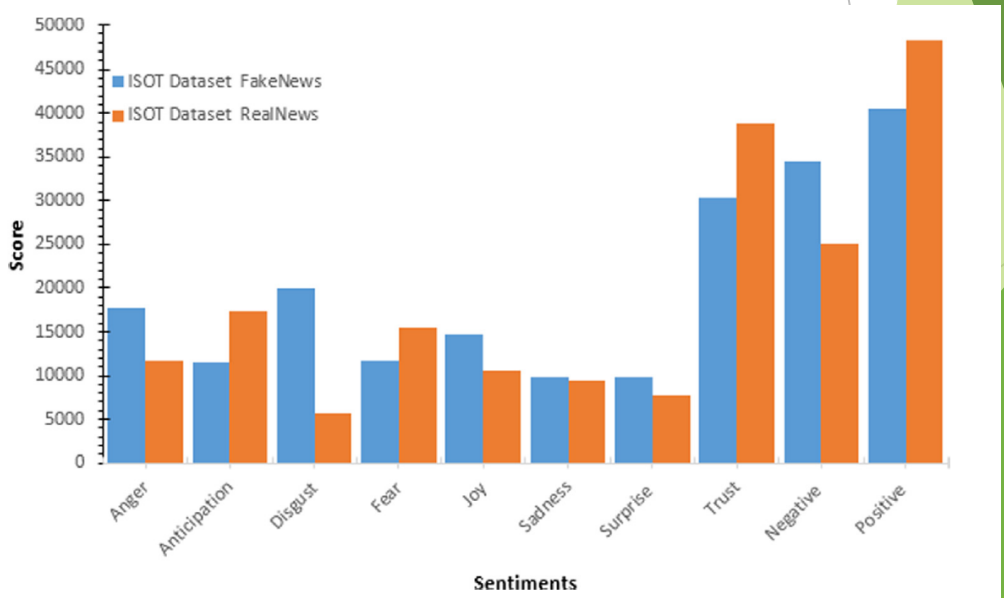



**Fig. 6.** ISOT Dataset fake and real news emotions.

# 2.Data Visualization

To gain insight into the datasets, writers carried out the visualization of datasets in the form word-cloud, pie charts, emotion graphs and frequency bar-graph. The purpose of visualization is to understand the structure of datasets.

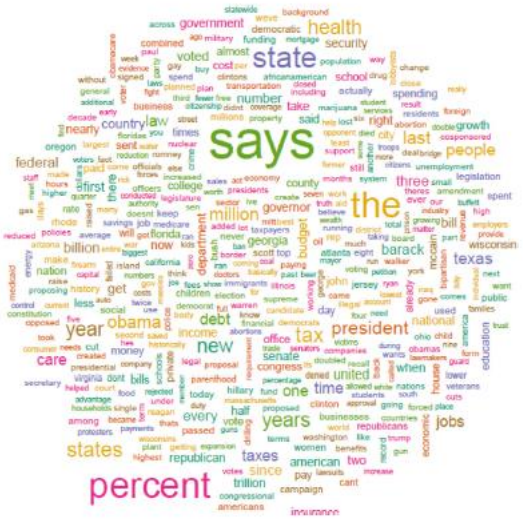



Fig. 3. Liar dataset real news word-cloud.

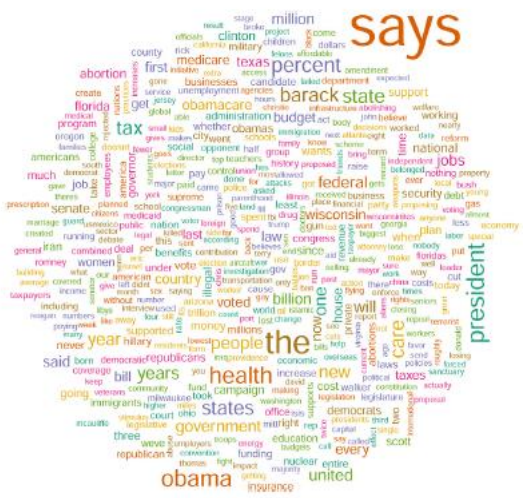



Fig. 4. Liar dataset fake news word-cloud.

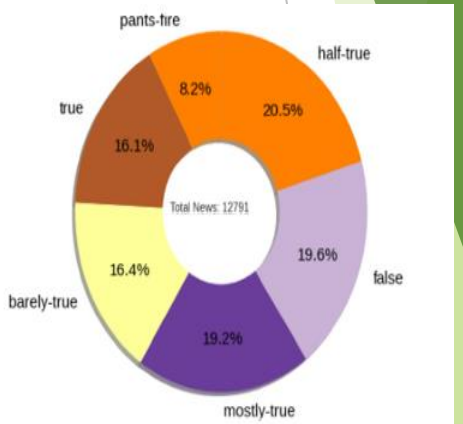



Fig. 9. Liar dataset class distribution.

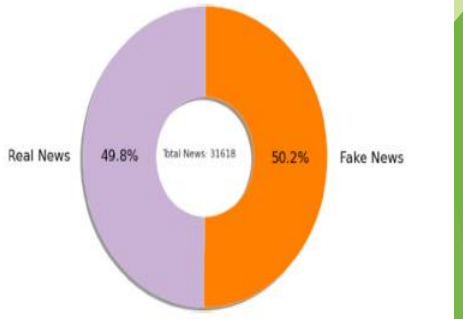



Fig. 10. ISOT Dataset class distribution.

# 3. Ensemble Model Selection

- An ensemble approach is a technique that blends the predictions of several machine learning-based algorithms to make more accurate predictions
- We identified and selected three popular supervised algorithms i.e. random forest, extra-tree algorithm and decision tree for the ensemble process.
- To aggregate the output of these multiple models, a bagging approach was used. The motivation for using the bagging method is stability that it provides to the model and also it reduces the possibility of overfitting in models.

# 4. Training of Model

- After following all the mentioned steps, the datasets were divided into training set and testing set using the k-fold approach.
- And train the model using training dataset.

# Experiment results

## 1.Experiment results for Liar dataset

| Classifier | Label | Before features extraction | | | After features extraction | | |
|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1-Measure | Precision | Recall | F1-Measure |
| Decision tree | barely-true | 17 | 14 | 16 | 38 | 39 | 40 |
| | false | 22 | 27 | 24 | 43 | 48 | 45 |
| | half-true | 24 | 21 | 22 | 40 | 44 | 42 |
| | mostly-true | 22 | 24 | 23 | 39 | 48 | 43 |
| | pants-fire | 20 | 15 | 17 | 51 | 40 | 45 |
| | true | 20 | 22 | 21 | 48 | 27 | 35 |
| Random forest | barely-true | 25 | 15 | 19 | 39 | 39 | 39 |
| | false | 24 | 49 | 32 | 44 | 49 | 46 |
| | half-true | 26 | 20 | 22 | 41 | 46 | 43 |
| | mostly-true | 27 | 29 | 31 | 40 | 50 | 44 |
| | pants-fire | 31 | 11 | 16 | 54 | 43 | 48 |
| | true | 28 | 19 | 23 | 41 | 27 | 35 |
| Extra tree classifier | barely-true | 24 | 16 | 19 | 36 | 37 | 36 |
| | false | 24 | 38 | 30 | 44 | 46 | 45 |
| | half-true | 26 | 25 | 25 | 41 | 44 | 43 |
| | mostly-true | 26 | 29 | 27 | 40 | 48 | 44 |
| | pants-fire | 31 | 13 | 19 | 45 | 41 | 43 |
| | true | 25 | 19 | 23 | 47 | 29 | 36 |

| Classifier | Accuracy and time complexity before feature extraction | | | | Accuracy and time complexity after feature extraction | | | |
|---|---|---|---|---|---|---|---|---|
| | Prediction accuracy | Prediction time (s) | Training accuracy | Training time (s) | Prediction accuracy | Prediction time (s) | Training accuracy | Training time (s) |
| Decision tree | 21.23 | 1.26 | 99.93 | 15.94 | 42.15 | 0.01 | 99.96 | 0.099 |
| Random forest | 25.92 | 1.42 | 99.93 | 25.66 | 44.15 | 0.40 | 99.96 | 1.70 |
| Extra tree classifier | 25.20 | 0.02 | 99.93 | 2.58 | 42.20 | 0.52 | 99.96 | 1.59 |

# 2. Experiment results for ISOT dataset

| Classifier | Label | Before features extraction | | | After features extraction | | |
|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1-Measure | Precision | Recall | F1-Measure |
| Decision tree | Fake | 100 | 99 | 100 | 100 | 100 | 100 |
| | Real | 99 | 100 | 99 | 100 | 100 | 100 |
| Random forest | Fake | 99 | 97 | 98 | 100 | 100 | 100 |
| | Real | 97 | 99 | 98 | 100 | 100 | 100 |
| Extra tree classifier | Fake | 97 | 98 | 98 | 100 | 100 | 100 |
| | Real | 97 | 98 | 98 | 100 | 100 | 100 |

| Classifier | Accuracy and time complexity before feature extraction | | | | Accuracy and time complexity after feature extraction | | | |
|---|---|---|---|---|---|---|---|---|
| | Prediction accuracy | Prediction time (s) | Training accuracy | Training time (s) | Prediction accuracy | Prediction time (s) | Training accuracy | Training time (s) |
| Decision tree | 99.29 | 0.05 | 100 | 5.36 | 100 | 0.01 | 100 | 0.32 |
| Random forest | 98.45 | 4.04 | 100 | 36.58 | 100 | 0.60 | 100 | 4.67 |
| Extra tree classifier | 97.59 | 4.78 | 100 | 65.09 | 100 | 0.01 | 100 | 0.32 |

# Application

- In daily life we can get the truthiness of any news or any information easily.
- For News Industry it is very helpful for detecting the fake news.
- Helpful for Stock Market where fake news are spread often.

# Conclusion

- In this article we have studied how we can deal with the spreading Of fake news using machine learning.
- We have studied a machine-learning based fake news detection model using a supervised approach.
- The experimentation of the model using liar and ISOT datasets yielded an accuracy of 44.15% and 100% percent.

# References:


[1] C. Buntain, J. Golbeck, Automatically identifying fake news in popular Twitter threads, in: 2017 IEEE International Conference on Smart Cloud, SmartCloud, IEEE, 2017, pp. 208– 215.

[2] B. Liu, J.D. Fraustino, Y. Jin, Social media use during disasters: A nationally representative field experiment, Tech. Rep., College Park, MD, 2013.

[3] Website: https://www.hindawi.com/journals/complexity/2020/8885861/ Fake News Detection using machine Learning

[4] YouTube Video : https://www.youtube.com/watch?v=tdFMIO5lfgA&ab_channel=GreatLearning Fake News Detection using machine Learning ,Great Learning,22July 2020.


**And Many More…**

# Questions?