Fake News Detection

### A Project Work Synopsis

*Submitted in the partial fulfillment for the award of the degree of*

# BACHELOR OF ENGINEERING

### IN

### Big DataAnalytics

### Submitted by:

### Prince Kumar

### 20BCS3936

### Rishab Kumar

### 20BCS3973

### Under the Supervision of:

### Sukhman Kaur



### CHANDIGARH UNIVERSITY, GHARUAN, MOHALI - 140413,

**PUNJAB**

#### OCTOBER 2022

# Table of Contents

Title Page Abstract

i iii

#### INTRODUCTION 4

* 1. Problem Definition 4
  2. Cause of problem 4

1. LITERATURE REVIEW 5
   1. Abstract of papers taken into consideration 5-6
   2. Literature Review Summary 7
2. PROBLEM FORMULATION 8-9
3. OBJECTIVES 9
4. BACKGROUND OF PROPOSED METHOD 10
5. **METHODOLOGY 11**
6. **RESULT AND DISCUSSION 12**
7. **CONCLUSION AND FUTURE SCOPE 13**
8. **REFRENCES 14**

**Abstract**

With the popularity of mobile technology and social media growing, information is readily available. Mobile App and social media platforms have overturned traditional media in the distribution of news. Alongside the increment in the utilization of online media stages like Facebook, Twitter, and so forth news spread quickly among a large number of clients with an extremely limited ability to focus time. Machine learning and Knowledge-based approach and approach are the two techniques utilized for investigating the truthiness of the content. Public and private assessments on a wide assortment of subjects are communicated and spread persistently through various online media. Most methodologies are utilized, for example, regulated AI. The spread of phony news has extensive results like the making of one-sided feelings to influencing political race results to support certain applicants. Additionally, spammers utilize engaging news features to produce income utilizing notices through click baits. In this paper, we intend to perform a parallel grouping of different news stories accessible online with the help of thoughts identifying with Artificial Intelligence, Natural Language Processing, and Machine Learning. The result of the project determines the fake news detection for social networks using machine learning and also checks the authenticity of the publishing news website.

# INTRODUCTION

#### 1.1 The growing popularity of social media & mobile technology with this information is accessible at one’s fingertips. Mobile apps and social media like Facebook and Twitter have overthrown traditional media in the field of information and news. With the convenience and speed that digital media offers, people express preference towards social media. Not only has it empowered consumers with faster access but it has additionally given benefit looking for parties a solid stage to catch a more extensive crowd.

#### With a lot of information or news, the one question occurred whether the given news or information is True or Fake. Fake news is commonly distributed with an intent to mislead or make an inclination to get political or monetary benefits. Let’s consider the example - In the recent elections of India, there has been a lot of discussion in regards to the credibility of different news reports preferring certain applicants and the political thought processes behind them. In this growing interest, exposing fake news is paramount in preventing its negative impact on people and society.

**1.2** The World Wide Web contains data in grouped arrangements like documents, videos, and audios. News distributed online in an unstructured configuration (like news, articles, videos, audios) is moderately hard to distinguish and order as this rigorously requires human mastery. However, computationalprocedures, for example, natural language preparing (NLP) can be utilized to identify irregularities that different a content article that is misleading in nature from articles that depend on realities. Different strategies include the investigation of the spread of fake news interestingly with real news. Specifically, this approach analyses fake news articles propagates differently on the internet relative to a true article. The reaction that an article gets can be separated at a theoretical level to arrange the article as real or fake. The hybrid approach can also be used to investigate the social responsibility of an article alongside investigating the text-based features to examine whether an article is deceptive or not.

The algorithms used by fake news detection systems include machine learning algorithms such as Logistic Regression, Random Forests, Decision trees, Support Vector Machines, Stochastic Gradient Descent, and so on. A simple method of fake news detection based on one of the AI algorithms called the Naive Bayes classifier help to examine how this particular method works for the particular problem with a manually labeled (fake or real) dataset and to support the idea of using machine learning to detect fake news

# LITERATURE REVIEW

[1] Paper Name: - Evaluating Machine Learning algorithms for Fake News Detection.

Author: - Shloka Gilda.

In this article, the author introduced the concept of the importance of NLP in stumbling across incorrect information. They have used time frequency-inverse document frequency (TF-IDF) of bigrams and probabilistic context-free grammar detection. Shloka Gilda introduced the concept of the importance of NLP in stumbling over incorrect information. They used Bi-Gram Count Vectorizer and Probabilistic Context-Free Grammar (PCFG) to detect deceptions. They examined the data set in more than one class of algorithms to find out a better model. The count vectorizer of bi-grams fed directly into a stochastic gradient descent model which identifies noncredible resources with an accuracy of 71.2%.

[2] Paper Name: - Fake News Detection on Social Media: A Data Mining Perspective.

Author: - Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang and Huan Liu.

In this paper to detect fake news on social media, a data mining perspective is presented that includes the characterization of fake news in psychology and social theories. This article looks at two main factors responsible for the widespread acceptance of fake messages by the user which is naive realism and confirmatory bias. It proposes a general two-phase data mining framework that includes 1) feature extraction and 2) modeling, analyzing data sets, and confusion matrix for detecting fake news.

[3] Paper Name: - Media Rich Fake News Detection: A Survey.

Author: - Shivam B. Parikh and Pradeep K. Atrey.

Social networking sites read news mainly in three ways: The (multilingual) text is analyzed with the help of computational linguistics, which semantically and systematically focuses on the creation of the text. Since most publications are in the form of text, a lot of work has been done on analyzing them. Multimedia: Several forms of media are integrated into a single post. This can include audio, video, images, and graphics. This is very attractive and attracts the viewer's attention without worrying about the text. Hyperlinks allow the author of the post to refer to various sources and thus gain the trust of viewers. In practice, references are made to other social media websites, and screenshots are inserted.

### Literature Review Summary

Table 2.1: Literature review summary

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Year and citation** | **Article Title** | **Purpose of the study** | **Comparison of technique done** | **Source (Journal/ Conference)** | **Findings** | **Data set (if used)** | **Evaluation parameters** |
| 2018  2017 | Media Rich Fake News Detection: A Survey.  Fake News Detection on Social Media: A Data Mining Perspective. | Origination of fake news   Application of existing algorithms through data mining perspective | social engagements on social media, and exploiting this auxiliary information  Deception Modeling, Predictive Modeling | Albany Lab for Privacy and Security  Cornell University | Different methods to detect different fake news  Fake news characterizations on psychology and social theories | BuzzFeedNews Dataset  Survey | Ngrams, Punctuation, Psycho-linguistic features  Evaluation metrics and representative datasets. |

# PROBLEM FORMULATION

# Social media facilitates the creation and sharing of information that uses computer-mediated technologies. This media changed the way groups of people interact and communicate. It allows low cost, simple access and fast dissemination of information to them. The majority of people search and consume news from social media rather than traditional news organizations these days. On one side, where social media have become a powerful source of information and bringing people together, on the other side it also 1 put a negative impact on society. Look at some examples herewith; Facebook Inc’s popular messaging service, WhatsApp became a political battle-platform in Brazil’s election. False rumours, manipulated photos, de-contextualized videos, and audio jokes were used for campaigning. These kinds of stuff went viral on the digital platform without monitoring their origin or reach.

# A nationwide block on major social media and messaging sites including Facebook and Instagram was done in Sri Lanka after multiple terrorist attacks in the year 2019. The government claimed that “false news reports” were circulating online. This is evident in the challenges the world's most powerful tech companies face in reducing the spread of misinformation. Such examples show that Social Media enables the widespread use of “fake news” as well

# disseminated on social media platforms may be of low quality carrying misleading information intentionally. This sacrifices the credibility of the information. Millions of news articles are being circulated every day on the Internet – how one can trust which is real and which is fake? Thus incredible or fake news is one of the biggest challenges in our digitally connected world. Fake news detection on social media has recently become an emerging research domain. The domain focuses on dealing with the sensitive issue of preventing the spread of fake news on social media.

# Fake news identification on social media faces several challenges. Firstly, it is difficult to collect fake news data.

# Furthermore, it is difficult to label fake news manually. Since they are intentionally written to mislead readers, it is difficult to detect them simply based on news content. Furthermore, Facebook, Whatsapp, and Twitter are closed messaging apps. The misinformation disseminated by trusted news outlets or their friends and family is therefore difficult to be considered as fake. It is not easy to verify the credibility of newly emerging and time-bound news as they are not sufficient to train the application dataset. Significant approaches to differentiate credible users, extract useful news features and develop authentic information dissemination systems are some useful domains of research and need further investigations. If we can’t control the spread of fake news, the trust in the system will collapse.

# OBJECTIVES

The proposed work is aimed to carry out work leading to the development of an approach for to solve fake news problem. The proposed aim will be achieved by dividing the work into following objectives:

1. Finding datasets containing both fake and real news alongwith their marker.

2.Model development and training on the datasets.

3.Model validation to improve the accuracy.

4.Creation of the web interface to deploy the model on the internet i.e.,

Create a Flask APP and a virtual environment.

5.Deployment of the model on the internet .

# 5 BACKGROUND OF PROPOSED METHOD

From an NLP perspective, researchers have studied numerous aspects of the credibility of online information. For example, [1] applied the time-sensitive supervised approach by relying on tweet content to address the credibility of a tweet in different situations. [2] used LSTM in a similar problem of early rumour detection. In another work, [3] aimed at detecting the stance of tweets and determining the veracity of the given rumour with convolution neural networks. A submission [4] to the SemEval 2016 Twitter Stance Detection task focuses on creating a bag-of-words autoencoder and training it over the tokenized tweets. Another team, [5], combined multiple models in an ensemble providing a 50/50 weighted average between a deep convolutional neural network and a gradient-boosted decision tree. Though this work seems to be similar to our work, the difference lies in the construction of an ensemble of classifiers. In a similar attempt, a team [6] concatenated various features vectors and passed them through an NLP model. Passive Aggressive algorithm is a margin-based online learning algorithm for binary classification. It is also an algorithm of a soft margin-based method and robust to noise. It can be used in fake news detection [16] Term Frequency-Inverse Document Frequency is also a method used to represent text in Fake News Detector 12 a format that can be easily processed by machine learning algorithms. It is a numerical statistic that shows how important a word is to news in a news dataset. The importance of a word is proportional to the number of times the word appears in the news (fake and real) but inversely proportional to the number of times the word appears in the news dataset (fake or real) [15]

# 6 METHODOLOGY

The following methodology will be followed to achieve the objectives defined for proposed research work:

1. Detailed study of existing system will be done.
2. Installation and hand on experience on existing approaches of proposed system will

be done. Relative pros and cons will be identified.

1. Various parameters will be identified to evaluate the proposed system.
2. Comparison of new implemented approach with exiting approaches will be done.

There are 3 main segments of the methodology :

* The core Machine Learning model
* The web interface.
* The common platform that brings the model and the interface together.

The Machine Learning Model There are two parts to the ML Model building. Machine Learning is a part of our life that can help us in predicting. We are using two types of model in this case. For the first part, we used passive-aggressive classifiers. And the steps include:

1. **Data** **Loading:** We are loading a CSV file for the data sorting and training-testing part of the model. The CSV file is turned into an array for easier work purpose.
2. **Vectorization:** Vectorization is needed for determining the frequency of the words present in a passage. This is needed to determine which words are used often.
3. **Classifier:** Passive-aggressive algorithms are a family of great learning algorithms. They are similar to Perceptron because it does not require a reading scale. However, unlike Perceptron, they include parameter correction. Passive is used when the prediction is correct and there is no change in the model. But if there is any kind of change in the model, that is if the prediction is not correct then the aggressive part is called, which changes the model accordingly. The aggressive part of the model changes the model according to its wish on the backend.
4. **Model Building:** The model is built through the train and test of the dataset, by ensuring that the training is done for 80% of the dataset and testing is done in the rest of the 20% of the dataset.

# 7 RESULTS AND DISCUSSION

In the fake news detection technology, there have been multiple instances where both unsupervised learning and supervised learning algorithms are used to classify text. Most of the literature survey focus on specific domains, most important the domain of politics. Therefore, the algorithm trained best works on a particular type of article’s domain and does not gives optimal results when presented to articles from different areas. Since articles from various areas have a special literary construction, it is hard to train a generic algorithm that works best on all specific news spaces. In this review paper, we find the solution for the fake news detection problem using the machine learning approach. We observed that the Random Forests algorithm with a simple term frequency-inverse document frequency vector gives the best output compares to others. Our study examines various text properties that can be used to distinguish fake and real content, and we trained a combination of different machine learning algorithms using these properties.

# 8 CONCLUSION AND FUTURE SCOPE

Manual classification of news articles requires in-depth knowledge and expertise in identifying anomalies in the text. It takes a lot of time to verify a single article manually that’s why we have discussed the problem of classifying fake news articles using machine learning models and ensemble techniques. It is important that we have a mechanism to detect fake news, or at least an awareness that not everything we read on social media may be true. That is why we always have to think critically. This way, we can help the people to make more informed decisions, and they won't be led to think about what others are trying to manipulate them into believing.

# 9 REFERENCES

[1] S. Gilda, "Notice of Violation of IEEE Publication Principles: Evaluating machine learning algorithms for fake news detection," 2017 IEEE 15th Student Conference on Research and Development (SCOReD), 2017, pp. 110-115, DOI: 10.1109/SCORED.2017.8305411.

[2] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), 2017, pp. 900-903, DOI: 10.1109/UKRCON.2017.8100379.

[3] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (n.d.)."fake news detection on social media: A data Mining Perspective".

[4] S. B. Parikh and P. K. Atrey, "Media-Rich Fake News Detection: A Survey," 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2018, pp. 436- 441, DOI: 10.1109/MIPR.2018.00093.

[5] C. Buntain and J. Golbeck, "Automatically Identifying Fake News in Popular Twitter Threads," 2017 IEEE International Conference on Smart Cloud (SmartCloud), 2017, pp. 208-215, DOI: 10.1109/SmartCloud.2017.40.

[6] A. Gupta and R. Kaushal, "Improving spam detection in Online Social Networks," 2015 International Conference on Cognitive Computing and Information Processing(CCIP), 2015, pp. 1-6, DOI: 10.1109/CCIP.2015.7100738.

[7] M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro, and L. de Alfaro, "Automatic Online Fake News detection Combining Content and Social Signals," 2018 22nd Conference of Open Innovations Association (FRUCT), 2018, pp. 272-279, DOI: 10.23919/FRUCT.2018.8468301.

[8] De Beer, Dylan, and Machdel Matthee. “Approaches to Identify Fake News: A Systematic Literature Review.” Integrated Science in Digital Age 2020 vol. 136 13–22. 5 May. 2020, doi:10.1007/978-3-030-49264-9\_2

[9] S. I. Manzoor, J. Singla, and Nikita, "Fake News Detection Using Machine Learning approaches A Systematic Review," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 230-234, DOI: 10.1109/ICOEI.2019.8862770.

[10] I. Ahmad, M. Yousaf, S. Yousaf, and M. O. Ahmad, “Fake News Detection Using Machine Learning Ensemble Methods,” Complexity, 17-Oct-2020. [Online]. Available: <https://www.hindawi.com/journals/complexity/2020/8885861/>.

[11] M. Gahirwal, “Fake News Detection,” International Journal of Advance Research, Ideas and Innovations in Technology, vol. 4, no. 1, pp. 817–819, 2018.

[12] Uma Sharma, Siddarth Saran, Shankar M. Patil, 2021, Fake News Detection using Machine Learning Algorithms, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NTASU – 2020 (Volume 09 – Issue 03).