

A Synopsis on

A Web Framework to Predict Fake News Using ML

Submitted in partial fulfillment of the requirements
of the degree of

Bachelor of Engineering

in

Information Technology

by

Jaagrut Shah (19204002)

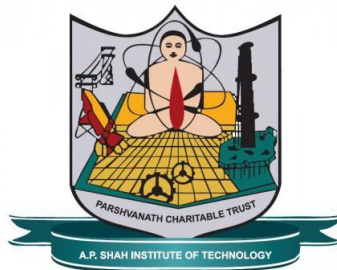
Jigar Desai (19204003)

Yash Jain (19204013)

Under the guidance of

Prof Apeksha Mohite

Prof Geetanjali Kalme



Department of Information Technology

A.P. Shah Institute of Technology
G.B.Road,Kasarvadavli, Thane(W), Mumbai-400615
UNIVERSITY OF MUMBAI
2020-2021

CERTIFICATE

This is to certify that the project Synopsis entitled “***A Web Framework to Predict Fake News Using ML***” Submitted by “***Jaagrut Shah (19204002), Jigar Desai (19204003), and Yash Jain (19204013)***” for the partial fulfillment of the requirement for the award of a degree ***Bachelor of Engineering in Information Technology.*** to the University of Mum-bai, is a bonafide work carried out during the academic year 2019-2020

Prof. Geetanjali Kalme
Co-Guide

Prof. Apeksha Mohite
Guide

Prof. Kiran Deshpande
Head Department of Information Technology

Dr. Uttam D.Kolekar
Principal

External Examiner(s)

1.

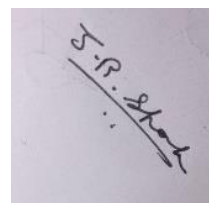
2.

Place: A.P.Shah Institute of Technology,Thane

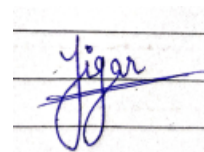
Date:

Declaration

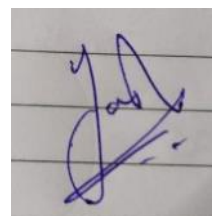
I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

A handwritten signature in black ink on a light-colored background. The signature is written in a cursive style and appears to read 'J. P. Shah'.

(Jaagrut Shah - 19204002)

A handwritten signature in blue ink on a light-colored background. The signature is written in a cursive style and appears to read 'Jigar'.

(Jigar Desai - 19204003)

A handwritten signature in blue ink on a light-colored background. The signature is written in a cursive style and appears to read 'Yash Jain'.

(Yash Jain - 19204013)

Date:12/11/2021

Abstract

Fake news has been a drag ever since the web boomed. The very network that allows us to know what is happening globally is the perfect breeding ground for malicious and fake news. Combating this fake news is vital because the world's view is formed by information. People not only make important decisions that support information but also form their own opinions. If this information is fake it can have devastating consequences. Verifying each news one by one by a person's being is unfeasible. This paper attempts to expedite the method of identification of faux news by proposing a system that will reliably classify fake news. Machine Learning algorithms such as Naive Bayes, Passive Aggressive Classifier, and Deep Neural Networks have been used on eight different datasets acquired from various sources. The paper also includes the analysis and results of every model. The arduous task of detection of faux news is often made trivial with the usage of the proper models with the proper tools. In the past years, the planet Wide Web (WWW) has become an enormous source of user-generated content and opinionative data. Using social media, such as Twitter, Facebook, etc, users share their views, feelings in a convenient way. Social media, such as Twitter, Facebook, etc, where millions of people express their views in their daily interaction, which can be their sentiments and opinions about particular things. These ever-growing subjective data are, undoubtedly, a particularly rich source of data for any quiet deciding process. To automate the analysis of such data, the world of Sentiment Analysis has emerged. It aims at identifying opinionative data within the Web and classifying them consistent with their polarity, i.e., whether or not they carry a positive or negative connotation. Sentiment Analysis may be a problem of text-based analysis, but some challenges make it difficult as compared to traditional text-based analysis. This clearly states that there is a need for an attempt to work towards these problems and it has opened up several opportunities for future research for handling negations, hidden sentiments identification, slangs, polysemy. However, the growing scale of knowledge demands automatic data analysis techniques.

In this paper, an in-depth survey on different techniques utilized in Sentiment Analysis is administered to know the extent of labor.

Introduction

The term “Fake News” was a lot less unheard of and not prevalent a couple of decades ago but in this digital era of social media, it has surfaced as a huge monster. Fake news, information bubbles, news manipulation, and therefore the lack of trust within the media are growing problems within our society. However, to start out addressing this problem, an in-depth understanding of faux news and its origins is required. Only then one can look into the different techniques and fields of Machine Learning (ML), Natural Language Processing (NLP), and Artificial Intelligence (AI) that could help us fight this situation. “Fake news” has been utilized in a mess of the way within the second half a year and multiple definitions are given. For instance, the NY Times defines it as “a made-up story to deceive”. Measuring fake news or even defining it properly could very quickly become a subjective matter, rather than an objective metric. In its purest form, fake news is made up, manipulated to resemble credible journalism and attract maximum attention, and, with it, advertising revenue. Despite these shortcomings, several entities have tried to categorize fake news in several manners.

Sentiment analysis which is also known as opinion mining is a method to an automatic finding of opinions incarnated in text, is becoming a challenge in many research areas, particularly in the data mining field for social media with several applications including product ratings and feedback analysis and customer decision making, etc [9]. Currently, social media has become a major public opinion finder and dissemination platform. With the expeditious development of Web 2.0, more and more people like to express their thoughts, views, and approaches over the Internet, which increases the vast source of user-generated content and opinionative data.

Objectives

- To show the news relevancy and analysis to attain accuracy in anticipating real and dependable news.

- To work on this issue, a layered model is proposed, which fine-tunes the information insight received from the data at each phase before attempting a prediction.
- To use a variety of Machine Learning approaches, achieve demonstratable success in the prediction of fake news and posts.
- To eliminate the propagation of false information on social media that may mislead users.
- To be able to give more and more accurate news on the screen.

Literature Review

The available literature has described many automatic detection techniques of fake news and deception posts. There are multidimensional aspects of fake news detection

ranging from using chatbots for the spread of misinformation to the use of clickbait for the spreading. There are many click baits available in social media networks including Facebook which enhance sharing and liking of posts which in turn spreads falsified information. A lot of work has been done to detect falsified information.

In this section, we discuss some of the papers that used machine learning to identify and classify fake news. Title: “A robust technique of fake news detection using Ensemble Voting Classifier and comparison with other classifiers”, Springer nature 2019. Author: Atik Mahabub Findings: Atik [7] uses a distinctive technique in detecting fake news by creating an 'Ensemble Voting classifier'. It uses many well-known Machine-Learning classifiers such as Naïve Bayes, Knn, SVM, and many more to verify the news. Further, cross-validation was used and the top three machine learning algorithms with the best accuracy were used in the Ensemble Voting Classifier. This model proposed a recognition structure that can productively predict the output and find the important highlights of the news. This allowed for a result ranging from the early to late 90s. Title: “Text-mining-based Fake News Detection Using Ensemble Methods”, International Journal of Automation and Computing. Author: Harita Reddy, Namratha Raj, Manali Gala, Annappa Basava. Findings: Text-mining-based methods for the detection of fake news have been evaluated by Harita Reddy et al [8]. This paper provided a hybrid approach that combines word vector representations and stylometric features using ensemble methods like bagging, boosting, and voting. After the selection of important features, Random Forest, Naïve Bayes, SVM, and many more algorithms were applied. This resulted from inaccuracies up to 95.49 %. Title: “Fake News Detection using Machine Learning and Natural Language Processing”, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-6. Author: Kushal Agarwalla, Shubham Nandan, Varun Anil Nair, D.Deva Hema Findings: The Natural Language Processing technique was exploited by Kushal Agarwalla et al [9]. to verify the news. NLTK from Python was used with various models including Logistic Regression, SVM, and Naïve Bayes with Lidstone Smoothing. Naive Bayes with 3

Lidstone smoothing performed admirably and gave a result of 83%. Perhaps, using only the vector-based methods to extract certain features and to train the classifiers is not an accurate solution as these are fixed to the particular dataset.

Title: “Fake News Detection using Naïve Bayes Classifier”, 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering. Author: Mykhailo Granik, Volodymyr Mesyura Findings: Mykhailo Granik et al [10]. implemented a basic approach using the Naive Bayes classifier for the detection of fake news. The model was built as a software system and validated over a set of Facebook news posts. It describes the similarity between spam messages and fake news articles by concluding that identical approaches can be taken for both fake news detection and spam filtering by producing a result of 75% accuracy.

Title: “Fake News Detection”, 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Sciences. Authors: Akshay Jain, Amey Kasbe Findings: Akshay Jain et al [11]. proposed a model with two variants that use the Naive Bayes classifier to predict whether a post on Facebook will be labeled as REAL or FAKE. The first model used the title as their source for vocabulary building, using a count vectorizer. And the second model used text as its source. The results were compared based on their AUC score and the second model was found to be better with a score of 0.93 and 0.912 with and without n_grams respectively.

Title: "Fake News Detection: A Deep Learning Approach," SMU Data Science Review: Vol. 1: No. 3, Article 10. Authors: Thota, Aswini; Tilak, Priyanka; Ahluwalia, Simrat; and Lohia, Nibrat (2018) Findings: Aswini Thota et al [2] used Deep Learning architectures to detect fake news. TfIDF, GloVe, and Word2Vec were used along with the DNN model to precisely predict the stance between the article body and given pair of the headline. This paper was able to produce an overall accuracy of 94.31%.

Title: ”The Pope Has a New Baby! Fake News Detection Using Deep Learning”, Stanford University CS 224N - Winter 2017. 4 Authors: Samir Bajaj Findings: Samir [12] explored different models ranging from Logistic Regression to CNN, RNN, and GRU. This work is mainly concentrated on using a pure NLP perspective to identify the presence of fake news by utilizing linguistic features. The highest precision of 0.97 was obtained using CNN with Max Pooling and Attention. This approach might

lose its viability as the fake news gets better at replicating true news. Title: “An Efficient Fake News Detection System Using Machine Learning”, International Journal of Innovative Technology and Exploring Engineering, Volume-8, Issue-10, August 2019. Authors: A. Lakshmanarao, Y. Swathi, T. Srinivasa Ravi Kiran,” Findings: A.Lakshmanarao et al [13]. employed SVM, Knn, Decision tree, and Random forest to build four models and compare them. It was observed that Random Forest Classification gave the highest score of 90.7% while least was provided by Support Vector Machines at 75.5%. Title: ”Evaluating Machine Learning Algorithms for Fake News Detection”, 2017 IEEE 15th Student Conference on Research and Development. Authors: Shlok Gilda, Findings: Shlok Gilda et al [14]. worked only using Natural Language Processing technique to identify the fake news. Probabilistic context-free grammar (PCFG) and Term frequency-inverse document frequency (TF-IDF) of bigrams were applied with various models like gradient boosting and stochastic gradient descent. Among other models, TFIDF of bi-grams with stochastic gradient descent identified fake news with higher accuracy. Title: ”Automatic Detection of Fake News”, Association for Computational Linguistics August 2018. Authors: Veronica Perez-Rosas, Bennett Kleinberg, Alexandra Lefevre, Rada Mihalcea, Findings: While the previous papers applied machine learning to the detection of fake news, Veronica [15] brings something new to the table in the form of human testing. This is the only paper that has tallied and compared the performance of humans against machines. This paper uses two different datasets namely FakeNewsAMT which consists of general news and Celebrity news which as the name suggests contains news about celebrities. The paper used two annotators to classify if the news were true or fake and they had an agreement rate of 70 5 percent. It is observed that the annotators beat the automated system when dealing with celebrity news datasets but lost by a margin of 3-4% in FakenewsAMT. Multi-feature extraction was also done and it showed that FakeNewsAMT performs best when relying on stylistic features and Celebrity on LIWC features.

Problem Definition

Social media for news consumption is a double-edged sword. On the one hand, its low cost, easy access, and rapid dissemination of information lead people to seek out and consume news from social media. On the other hand, it enables the widespread of “Fake News”, i.e., low-quality news with intentionally false information. The extensive spread of fake news has the potential for extremely negative impacts on individuals and society. Therefore, fake news detection on social media has recently become emerging research that is attracting tremendous attention. Fake news detection on social media presents unique characteristics and challenges that make existing detection algorithms from traditional news media ineffective or not applicable. First, fake news is intentionally

written to mislead readers to believe false information, which makes it difficult and nontrivial to detect based on news content; therefore, we need to include auxiliary information, such as using social engagements on social media, to help decide. Second, exploiting this auxiliary information is challenging in and of itself as users' social engagements with fake news produce data that is big, incomplete, unstructured, and noisy.

Proposed System Architecture/Working

Existing System:

Various researchers have attempted to solve this challenge in a multitude of ways to test which method works and get desirable results. A few studies have discussed fake news detection approaches from a data mining perspective, including feature extraction and model construction. A methodology of feature extraction (both news content features and social context features) combined with metric evaluation using precision, recall, and f1 scores has proved to bear educated results but the problem is not that simple. Other parameters like bot spamming, clickbait, source of news also affect the predictions [1]. These were some data mining and NLP-oriented approaches, but with more and more research and development in AI, researchers became interested in heavy neural network-oriented approaches. A paper showed a method of 'capture', 'score', 'integrate' and

creates a model of recurrent neural networks for stance detection of fake news. They used a recurrent neural network to capture the temporal pattern of user activity around a specific article/text, then the user behavior is used to extract source characteristics. All this information is used and integrated to form a model for the classification of fake news [3]. The studies prove that even a simple straightforward network model can outperform complex models. So clearly, the complexity of the model is not the optimum solution here and rather the right choice of parameters and data is essential. Even the use of convolutional neural networks has been done while some others tried linguistically infused neural networks. Another paper discusses linguistically-infused neural network model with Long-Short-Term-Memory (LSTM) and Convolutional Neural Network (CNN) to classify Twitter posts. The linguistic part was introduced using the GloVe library of pre-trained vectors [4]. So, it is evident that many attempts have been made but it is all a bit messy and scattered. There is a lot of room for development and research in this area especially because news statements have so many variables attached to them: sarcasm, abbreviation, metaphors, etc. However, efforts have been made to arrange reliable and vast data into a quality dataset. One such benchmark dataset has been used in this project. The fake news problem is growing at an alarming rate and it needs to be addressed more aggressively.

A. Data:

Categorizing a news statement as “fake news” could be a very challenging and time-consuming task. For this reason, the use of an existing dataset, that has already collected and classified fake news, has been made. The data source used for this 7 project is the LIAR dataset. Given below is a brief description of the data files used for this study. The dataset has been cited in the paper [5] "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection, to appear in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), short paper, Vancouver, BC, Canada, July 30-August 4, ACL [5]. The original dataset contained 13 variables/columns for train, test, and validation sets. For the sake of simplicity, only two variables from this original dataset have been chosen for this classification task. The other

variables may be used later to achieve a more detailed analysis. The two columns that have been used are ‘Statement’, which is the actual news statement itself, and ‘Label’, which refers to the statement being true/false. The procedure used for reducing the number of classes in the dataset: True, mostly true, half-true become ‘True’. Barely true, false, pants-fire becomes ‘False’. An important point to note here is that this dataset comes with 3 separate TSV files for train, validation, and test. So, it is not required to manually split the data into test, train, validation.

B. Model:

The performance of a classifier may vary based on the size and quality of the text data (or corpus) and also the features of the test vectors. Common noisy words called ‘stop words’ are less important words when it comes to text feature extraction, they don’t contribute towards the actual meaning of a sentence and they only contribute towards feature dimensionality and may be discarded for better performance. This helps in reducing the size/dimensionality of the text corpus and adding text context for feature extraction. Also, lemmatization is used to convert words to their core meaning, and this results in multiple word conversion into a single discrete representation [6].

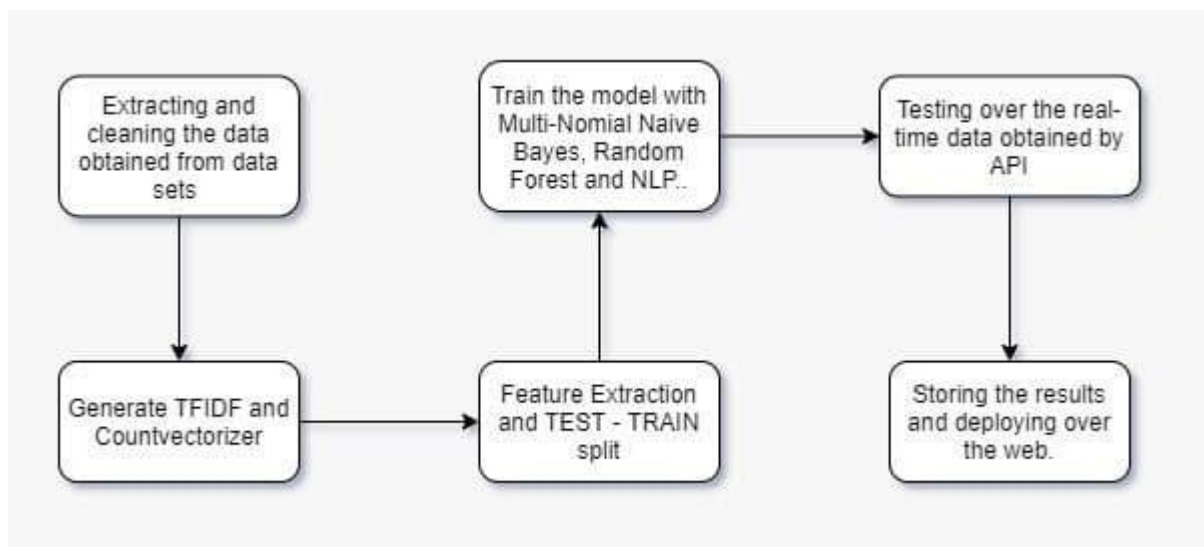


Fig no1: Block Diagram

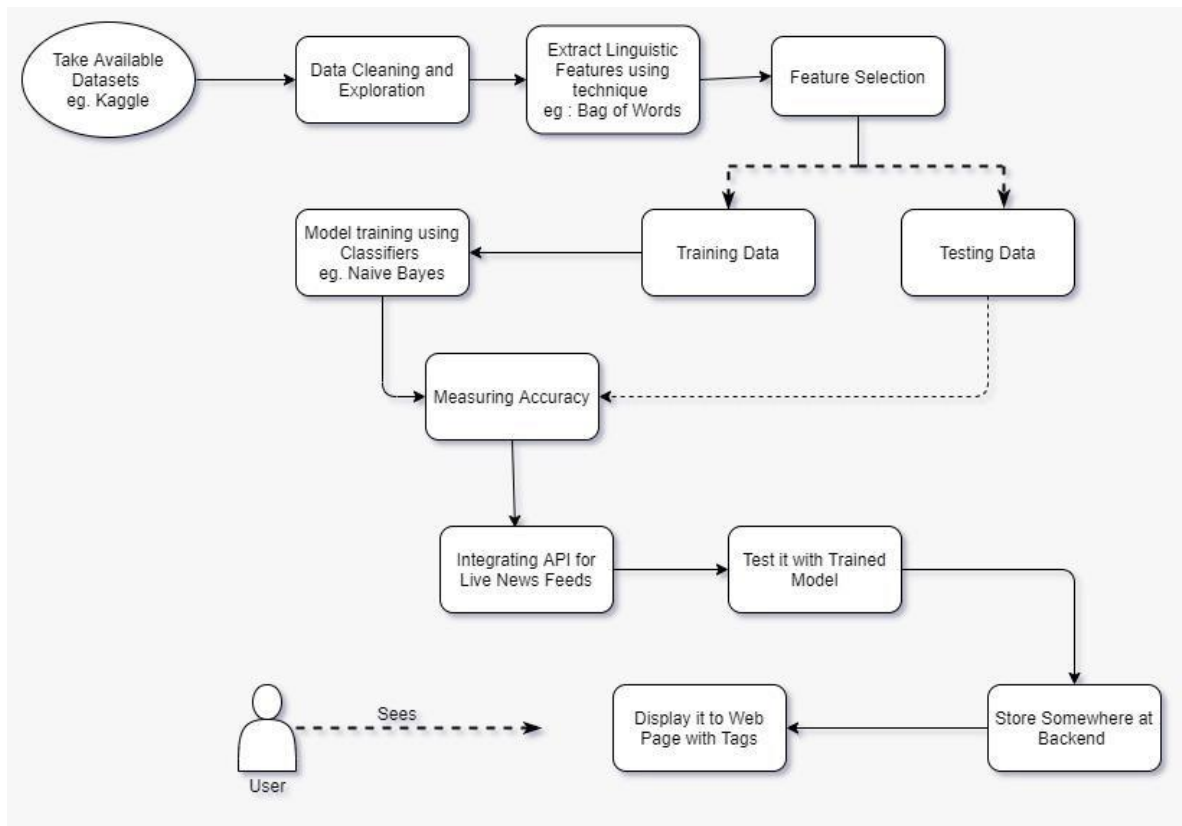


Fig no2: Flow Diagram

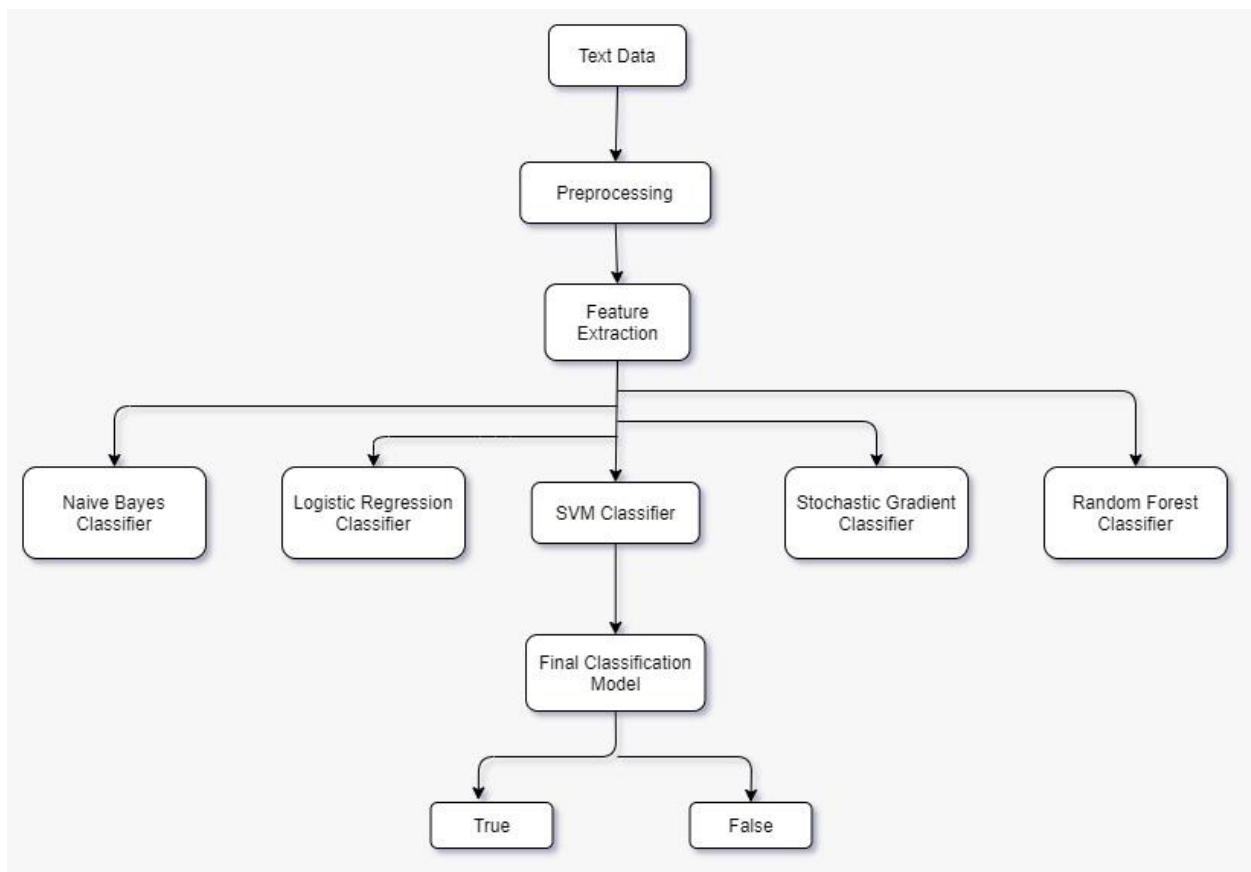


Fig no3: Proposed System

Summary

The fake news challenge is perilous and is spreading rapidly like a wildfire as it becomes easier for information to reach the mass in various flavors. Reports have shown that, just like in the last US presidential elections, fake news can have a huge impact in politics and thereafter on the people like a domino effect. With the help of Machine Learning & Artificial Intelligence, we can control and limit the spread of such misinformation more quickly and efficiently as compared to manual efforts. The work in this project proposes a stacked model which fine-tunes the informational insight gained from the data at each step and then tries to make a prediction. Machine Learning has opened a new front in warfare against Fake news, one must take advantage of this front and exploit it thoroughly. This report has shown that the front is viable. The usage of Machine learning in the identification of fake news is still in its infancy. Every model built or a system proposed is one step closer to a fake news-free internet

References

- [1] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). "Fake news detection on social media: A data mining perspective" . *ACM SIGKDD Explorations Newsletter*, 19(1), 22-36.
- [2] Thota, Aswini; Tilak, Priyanka; Ahluwalia, Simrat; and Lohia, Nibrat (2018) "Fake News Detection: A Deep Learning Approach," *SMU Data Science Review*: Vol. 1: No. 3, Article 10.
- [3] Ruchansky, N., Seo, S., & Liu, Y. (2017, November). "CSI: A hybrid deep model for fake news detection" . In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 797-806). ACM.
- [4] Volkova, S., Shaffer, K., Jang, J. Y., & Hodas, N. (2017, July). "Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter" . In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 647-653).
- [5]. Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*
- [6]. Reis, J. C., Correia, A., Murai, F., Veloso, A., Benevenuto, F., & Cambria, E. (2019). "Supervised Learning for Fake News Detection" . *IEEE Intelligent Systems*, 34(2), 76-81.
- [7]. Atik Mahabub, "A robust technique of fake news detection using Ensemble Voting Classifier and comparison with other classifiers" , Springer nature 2019.
- [8]. Harita Reddy, Namratha Raj, Manali Gala, Annappa Basava, "Text-mining-based Fake News Detection Using Ensemble Methods" , *International Journal of Automation and Computing*.
- [9]. Kushal Agarwalla, Shubham Nandan, Varun Anil Nair, D.Deva Hema, "Fake News Detection using Machine Learning and Natural Language Processing" , *International Journal of Recent Technology and Engineering (IJRTE)* ISSN: 2277-3878, Volume-7, Issue-6.
- [10]. Mykhailo Granik, Volodymyr Mesyura, "Fake News Detection using Naïve Bayes Classifier" , 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering.
- [11]. Akshay Jain, Amey Kasbe, "Fake News Detection" , 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Sciences.
- [12]. Samir Bajaj, "The Pope Has a New Baby! Fake News Detection Using Deep Learning" . Stanford University CS 224N - Winter 2017.
- [13]. A. Lakshmanarao, Y. Swathi, T. Srinivasa Ravi Kiran, "An Efficient Fake News Detection System Using Machine Learning" , *International Journal of Innovative Technology and Exploring Engineering*, Volume-8, Issue-10, August 2019.
- [14]. Shlok Gilda, "Evaluating Machine Learning Algorithms for Fake News Detection" , 2017 IEEE 15th Student Conference on Research and Development.
- [15]. Veronica Perez-Rosas, Bennett Kleinberg, Alexandra Lefevre, Rada Mihalcea, "Automatic Detection of Fake News" , *Association for Computational Linguistics* August 2018.
- [16]. Ammara Habib, Muhammad Zubair Asghar, Adil Khan, Anam Habib, Aurangzeb Khan, "False information detection in online content and its role in decision making: a systematic
- [17]. Kaggle, Fake News, Kaggle, San Francisco, CA, USA, 2018, <https://www.kaggle.com/c/fake-news>.
- [18]. Kaggle, Fake News Detection, " " San Francisco, CA, USA, 2018, <https://www.kaggle.com/jruvika/fake-news-detect>

Publication

2nd International Conference on Advance Computing And innovation Technologies in
Engineering
Greater Noida, India