A Project Report on

Fake News Analysis

Submitted in partial fulfilment of the requirements for the award of the degree of

Bachelor of Engineering

in

Computer Engineering

by

Shubham Padte (16102058) Ashish Kothari (16102045) Pankit Khimasiya (16102034)

Under the Guidance of

Dr. Rahul Ambekar

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

Academic Year 2019-2020

Approval Sheet

This Project Report entitled *"Fake News Analysis"* Submitted by *"Shubham Padte "(16102058) "Ashish Kothari"(16102045),"Pankit Khimasiya"(16102034)*is approved for the partial fulfilment of the requirement for the award of the degree of *Bachelor of Engineering* in *Branch Name* from *University of Mumbai*.

(Dr.Rahul Ambekar)
Guide

Prof. Sachin Malve
Head Department of Computer Engineering

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

CERTIFICATE

This is to certify that the project entitled *"Title of project"* submitted by *"Shubham Padte"(16102058),"Ashish Kothari" (16102045),"Pankit Khimasiya" (16102034)* for the partial fulfilment of the requirement for award of a degree *Bachelor of Engineering* in *Computer Engineering* ,to the University of Mumbai, is a bonafi de work carried out during academic year 2019-2020.

(Dr.Rahul Ambikar)
Guide

Prof. Sachin Malve
Head Department of Computer Engineering

Dr. Uttam D.Kolekar
Principal

External Examiner(s)

1.

2.

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

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that We have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We 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.

——————————————————

(Signature)

——————————————————

(Shubham Padte 16102058)
(Ashish Kothari 16102045)
(Pankit Khimasiya 16102034)

Date:

Abstract

Social media for news consumption is a double-edged sword. A large body of recent works has focused on understanding and detecting fake news stories that are disseminated on social media. To accomplish this goal, these works explore several types of features extracted from news stories, including source and posts from social media. Extensive spread of fake news has potential for extremely negative impacts on individuals and society. Fake news is intentionally written to mislead readers to believe false information. Proposed application will be implemented to verify the gentility of news. We believe that interesting findings on usefulness and importance of features for detecting false news will prevent readers from misleading. This project will filter news by using Natural language processing and Machine learning, by processing large dataset of news.

Contents

List of Abbreviations

ACK:	Acknowledgement
NLP:	Natural Language Processing
W2Vec:	Word to vector
Tf-idf:	Term frequency inverse document frequency

Chapter 1

Project Conception and Initiation

1.1 Introduction

Internet have been dramatically changing the way news is produced, disseminated, and consumed, opening unforeseen opportunities, but also creating complex challenges. A key problem today is that social media has become a place for campaigns of misinformation that affect the credibility of the entire news ecosystem.

As an increasing amount of our lives is spent interacting online through social media platforms and other online platforms, more and more people tend to seek out and consume news from social media rather than traditional news organizations. The reasons for this change in consumption behaviours are inherent in the nature of these social media platforms: (i) It is often timelier and less expensive to consume news on social media compared with traditional news media, such as newspapers or television. (ii) It is easier to further share, comment on, and discuss the news with friends or the other reader on social media. Our economies are not immune to the spread of fake news either, with fake news being connected to stock market fluctuations and massive trades. For example, fake news claiming that Barack Obama was injured in an explosion wiped out 130 billion $in stock value

1.2 Objectives

This is the standalone application that will use the dataset which is consists of various information in mixture it contains fake news and real news and also the news that appear real but are fake The major objective is detect fake news , which is a classic text classification problem with a straight forward proposition. It is needed to build a model that can classify “fake” or “real” news. To deliver user our opinion on news and make them classify the news is fake or real.

1.2.1 Problem Definition

To Design and Develop a Model for Fake news that would overcome the problems created due to false news spread on internet. We propose a general data mining framework for fake news detection which includes two phases: (i) feature extraction and (ii) model construction Our proposed application will be implemented to verify the gentility of news.

1.2.2 Scope

A model will be trained by letting it read millions of documents. These documents included journalism and scientific journal articles, satire articles, narrative fiction, opinion pieces, blogs, politically leaning news and even hate speech examples. The model still learning all the time. If you think the model’s opinion wasn’t quite right then we will include that feedback as a training data set for the model which will continue to learn and develop using the new data. There are many types of false news, ranging from satire to propaganda. In this model, we focus on text-only documents formatted as news articles: stories and their corresponding metadata that contain purposefully false information. Existing fake news is predominantly human-written.

1.2.3 Technology Stack

- **Jupyter notebook**: It is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning etc.

- **Anaconda:** Anaconda is a free and open source distribution of the Python and R programming languages for data science and machine learning related applications (large-scale data processing, predictive analytics, scientific computing), that aims to

simplify package management and deployment. Package versions are managed by the package management system anaconda.

- **Python:** Python is an interpreted, object-oriented, high level programming with dynamic semantics. Its high-level built-in data structures, combined with dynamic typing and binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance.

- **Machine learning:** Machine learning explores the construction of algorithms which can learn and make predictions on data. Such algorithms follow programmed instructions, but can also make predictions or decisions based on data. They build a model from sample inputs.

- **Naive Bayes Algorithm:** In machine learning, naive Bayes classifiers are a family of simple" probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

- **Deep Learning:** Deep learning is part of a of machine learning methods based on learning data representations, as opposed to task-specific algorithms. Learning can be supervised, semi-supervised or unsupervised. Deep learning architectures such as deep neural networks, deep belief networks and recurrent neural networks have been applied to fields including computer vision, speech recognition, processing, social network filtering, machine translation, bioinformatics, drug design and board game programs, where they have produced results comparable to and in some cases even exceeded the human experts.

1.2.4 Benefits for Society

Correcting the fake news is one of the most important aspects for the society. Main goal of this project is to prevent the society for being misguided as well as spreading the rumours. By reducing the spread of rumours there can be prevention of chaos in people.

Chapter 2

Literature Review

Supervised Learning for Fake News Detection Julio C. S. Reis, Andre Correia, Fabricio Murai, Adriano Veloso, and Fabricio Benevenuto Universidade Federal de Minas Gerais.

This paper briefly surveys existing studies on this topic, identifying the main features proposed for this task. We implement these features and test the effectiveness of a variety of supervised learning classifiers when distinguishing fake from real stories on a large, recently released and fully labelled dataset. Finally, we discuss how supervised learning models can be used to assist fact-checkers in evaluating digital content and reaching warranted conclusions.

Fake News Detection on Social Media: A Data Mining Perspective Kai Shu†, Amy Sliva‡, Suhang Wang†, Jiliang Tang and Huan Liu† †Computer Science Engineering, Arizona State University, Tempe, AZ, USA ‡Charles River Analytics, Cambridge, MA, USA.

In this paper we propose a general data mining framework for fake news detection which includes two phases: (i) feature extraction (by types of news content and social context features) (ii) model construction (by types of news content and social context features). We also discussed the datasets and evaluation metrics used by existing methods.

Fake News Detection Akshay Jain, Amey Kasbe Department of Electronics and Communication Engineering Maulana Azad National Institute of Technology Bhopal, India.

This paper describes a simple fake news detection method based on one of the machine learning algorithms – na¨ıve Bayes classifier. The goal of the research is to examine how na¨ıve Bayes works for this particular problem, given a manually labelled news dataset, and to support (or not) the idea of using artificial intelligence for fake news detection

Chapter 3

Project Design

3.0.1 Proposed System

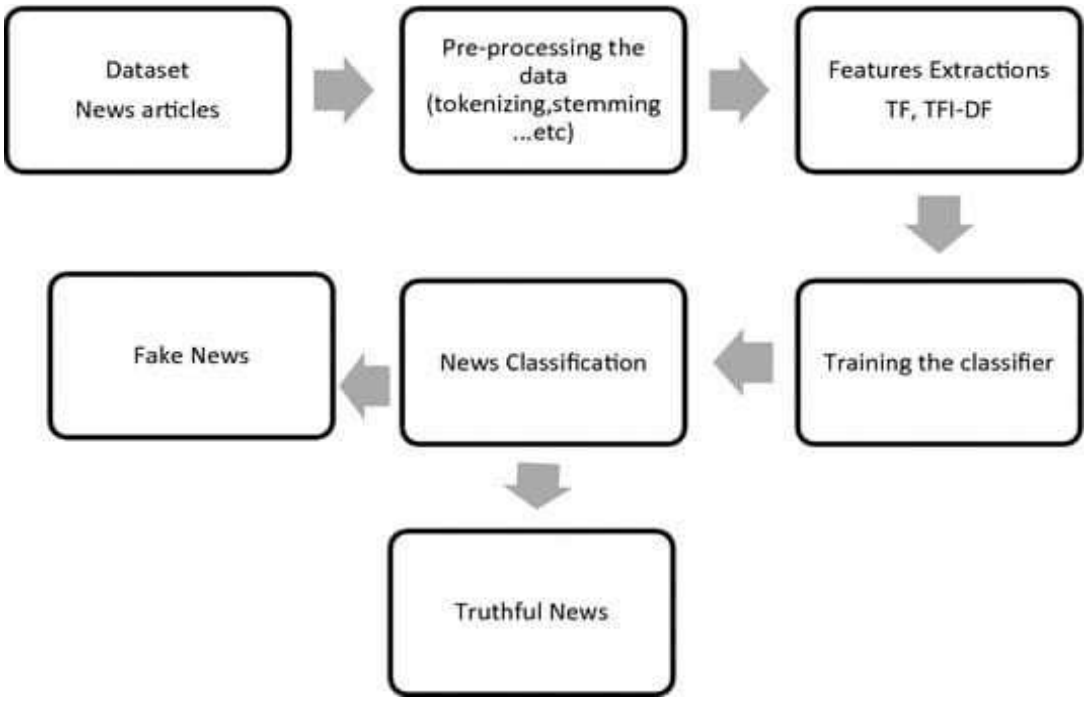
1. In this paper a model is build based on the count vectorizer or a tf-idf matrix i.e. word tallies relatives to how often they are used in other articles in your data set can help.

2. Since this problem is a kind of text classification, implementing a Naive Bayes classifier will be best as this is standard for text-based processing.

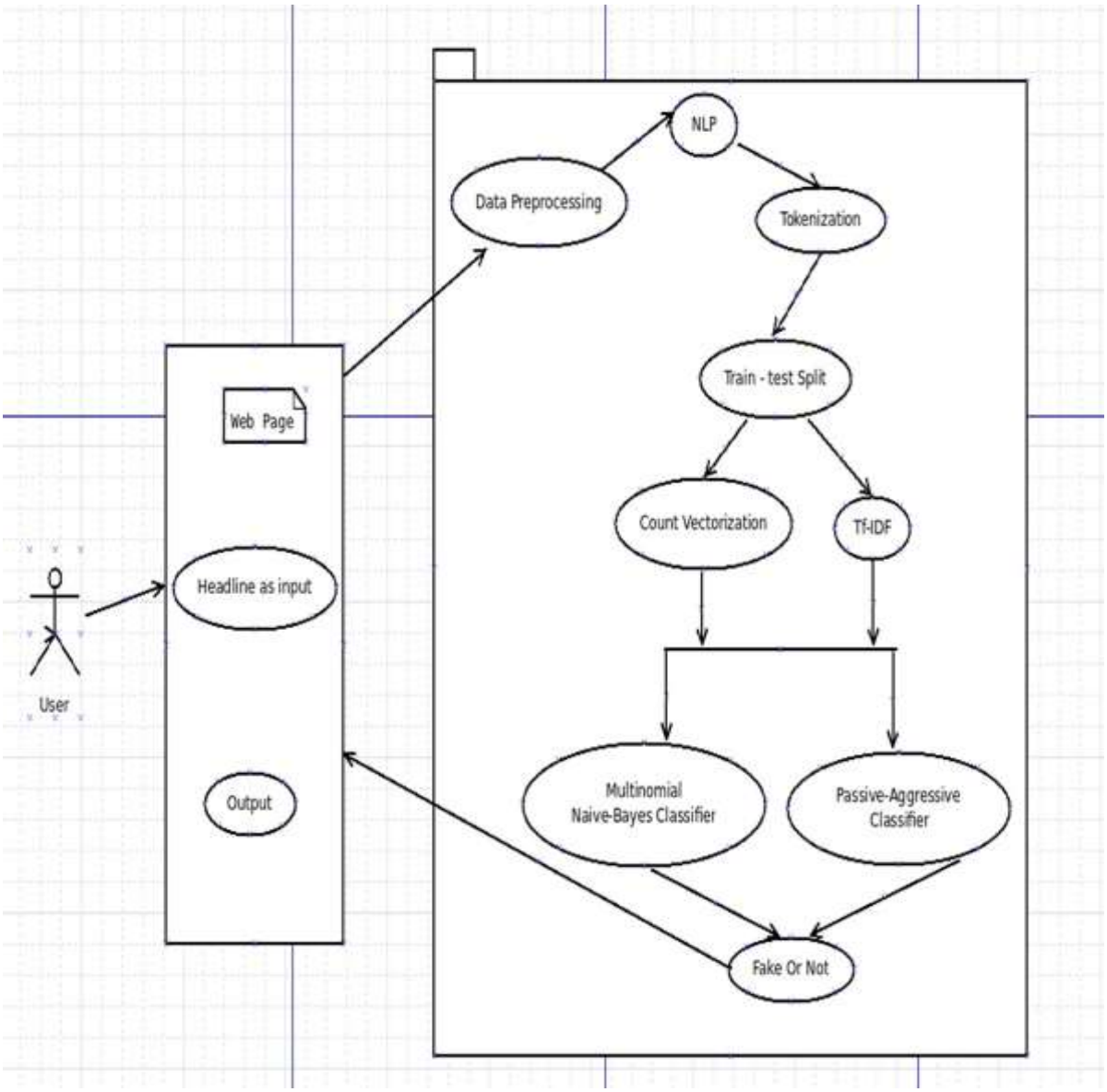
3. The actual goal is in developing a model which was the text transformation (count vectorizer vs tf-idf vectorizer) and choosing which type of text to use (headlines vs full text).

4. Now the next step is to extract the most optimal features for count vectorizer or tf-idf vectorizer, this is done by using a n-number of the most used words, and/or phrases, lower casing or not, mainly removing the stop words which are common words such as “the”, “when”, and “there” and only using those words that appear at least a given number of times in a given text data-set.

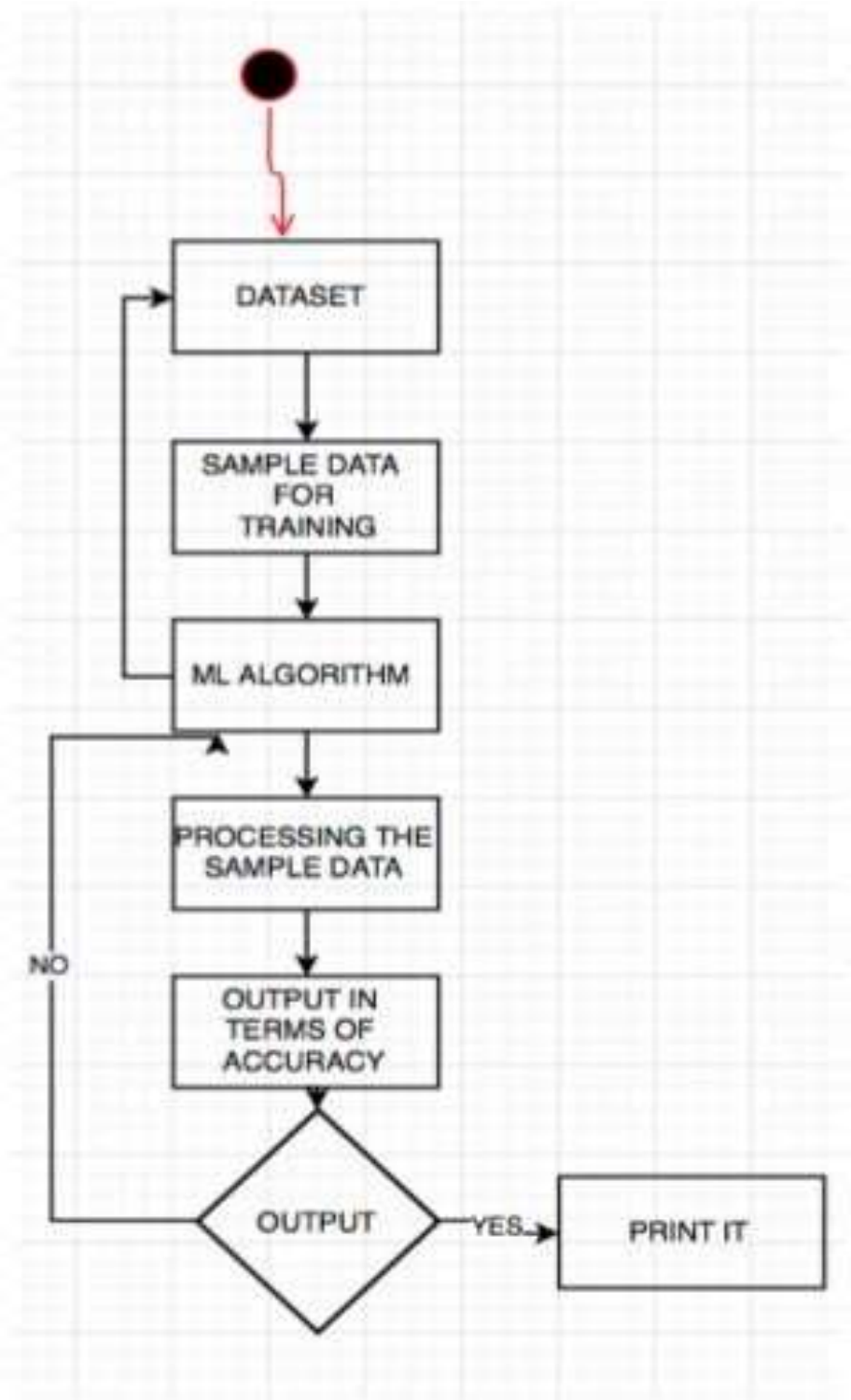
3.0.2 Flow of Modules



3.0.3 Use Case Diagram



3.0.4 Activity Diagram



Chapter 4

Project Implementation

4.0.1 Module 1

Dataset: Collecting the dataset for training and testing is the most challenging part of this project. The dataset will be divided into two parts namely fake and real. Now fake news dataset we can get from Kaggle as fake news is already fake. But collecting the real news dataset is the most challenging. We will collect the real news dataset from numerous articles and verify it thoroughly whether it's actually real or not. Because the real news dataset will widely affect our accuracy and will help the algorithm to understand better. **Pre-processing:** Pre-processing includes removal of stop words like What, If, Then, And, The etc. Basically, words which do not contribute in the algorithm. Lemmatizing the word which means grouping of same word in different tenses and forms for e.g.: played, playing, playful will be in one group and treated as same words. We also do normalization which includes converting text into more uniform and standard format. Normalization is essential before we use passive aggressive classifier.

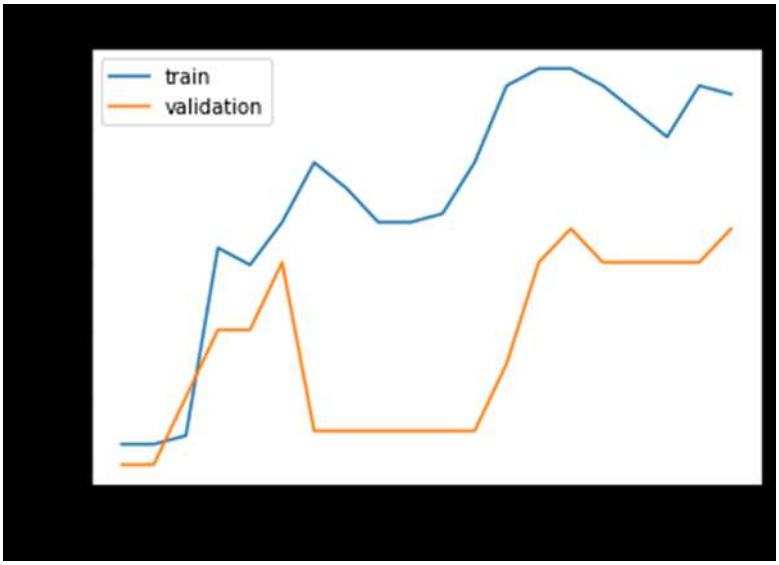
4.0.2 Module 2

Algorithms: After pre-processing the dataset, we will apply TFIDF and count vectorizer. TFIDF Count vectorizer: Gives us the count of a particular word appearing in a document and also the importance of that word in the document. Then we will split the dataset into training and testing using Train-Test Split. Then we apply Multinomial Naive Bayes and Passive Aggressive Classifier.
Multinomial Naive Bayes: This algorithm will calculate the probability of a word with respect to the document. It will help us classify the text according to their context Passive Aggressive Classifier: It is an online algorithm. We cannot keep each and every label of numerous articles and documents in a memory, so we need an algorithm which gets an example learns from it and throws it away.

Chapter 5

Results

These results directly induce the accuracy percentage of the given headline in a dataset which are trained and tested by our model. The results also show the graph of per given headlines by the input given by the user. How much valid is the given headline in the input can be verified from the received output graph or the accuracy level.



Chapter 6

Conclusions and Future Scope

This shows a simple approach for fake news detection using Classification and NLP methods. This approach was implemented as a software system and tested against a data set of news and some social media posts. Thus, this system will be helpful to users to recognize that which headline is fake which one is real. .

Bibliography

[1] Julio C. S. Reis, Andre Correia, Fabricio Murai, Adriano Veloso, and Fabricio Benevenuto "Supervised Learning for Fake News Detection" IEEE research paper 2019.[1]

[2] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu "Fake News Detection on Social Media: A Data Mining Perspective" IEEE research paper Sep 2017.[2]

[3] Akshay Jain, Amey Kasbe "Fake News Detection" 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Sciences[3]

Appendices

Appendix-A: Registration on BBC news/Kaggle console

1. Go to the site' https://console.BBCnews.com '
2. Register with google account to make use of free services.
3. Logout from the account.

Appendix-B:

1. Registration on Fakerfact.com
2. Blogs of ALT NEWS

Acknowledgement

We have great pleasure in presenting the report on **Fake News Analysis.** We take this opportunity to express our sincere thanks towards our guide **Dr.Rahul Ambikar** Department of Computer Engineering, APSIT thane for providing the technical guidelines and suggestions regarding line of work. We would like to express our gratitude towards his constant encouragement, support and guidance through the development of project.

We thank **Prof. Sachin Malve** Head of Department,Computer Engineering, APSIT for his encouragement during progress meeting and providing guidelines to write this report.

We thank **Prof.Amol Kalugade** BE project co-ordinator, Department of Computer Engineering, APSIT for being encouraging throughout the course and for guidance.

We also thank the entire staff of APSIT for their invaluable help rendered during the course of this work. We wish to express our deep gratitude towards all our colleagues of APSIT for their encouragement.

Student Name1:Shubham Padte
Student ID1:16102058

Student Name2:Ashish Kothari
Student ID2:16102045

Student Name3:Pankit Khimasiya
Student ID3:16102034

Publication

Paper entitled **"Fake News Detection "** is presented at **"2018 IEEE International Students' Conference on Electrical, Electronics and Computer Sciences "** by **"Akshay Jain".**