FAKE NEWS DETECTION

A Report for of Capstone Design -II

S.No	Enrollment Number	Admission Number	Student Name	Degree / Branch	Sem
1	1713101689	17SCSE101722	ANJALI SINGH	Btech/CSE	VIII
2	1713101289	17SCSE101299	HITENDRA DHAKAREY	Btech/CSE	VIII
3	1713101316	17SCSE101326	TANMAY SHUKLA	Btech/CSE	VIII

Under the Supervision of

Ms Heena Khera, Assistant Professor

School of Computing Science and Engineering

Greater Noida, Uttar Pradesh

Fall 2020 - 2021

DECLARATION

We, the students of Galgotias University named Anjali Singh, Tanmay Shukla and Hitendra Dhakarey of BTech Computer Science of Sem-8 declare that the project report entitled “FAKE NEWS DETECTION” submitted by us is our own work and has been carried out under the supervision of Ms Heena Khera of Galgotias University.

We also declare that this project has not been previously submitted to any other university.

Date: 27/01/2021
Place:Greater Noida

ACKNOWLEDGEMENT

We three students of BTech Computer Science and Engineering of 4th year in Galgotias University are developing the final year project named “FAKE NEWS DETECTION”. We wholeheartedly express our sincere gratitude to Ms Heena Khera who guided us throughout for the final year project. We are also thankful to all the teachers for explaining the critical aspect related to this project. We would like to thank all the faculties who helped us and taught us the fundamentals which would lead to the completion of this project.

Anjali Singh

Tanmay Shukla

Hitendra Dhakarey

ABSTRACT

Fake news has quickly become a society problem, being used to propagate false or rumour information in order to change people's behaviour,this topic on social media has recently attracted tremendous attention. It has become a common sight to watch people fall prey to the fake news and end up taking actions which leads to grave consequences and affect the society as a whole.

This adversary needs to be stopped completely so that valuable information can be accessed without much of a hindrance. This problem can be easily solved if we use the technology around us and start using them as a means to stop the spread of the fake news.
The project entitled “FAKE NEWS DETECTION” has the sole purpose of solving this problem by identifying the news as fake or real by just a click is what the deliverable function of this project.

This project will act as a tool for the users to differentiate between a rumour and a fact. This way the spread and growth of the fake news will be controlled to some extent and valuable information will not be adulterated as well.

TABLE OF CONTENTS

Introduction About Project

The basic countermeasure of comparing websites against a list of labeled fake news sources is inflexible, and so a machine learning approach comes into picture.

This Project comes up with the applications of NLP (Natural Language Processing) techniques for detecting the 'fake news', that is, misleading news stories that come from non-reputable sources. Combatting the fake news is a classic text classification project with a straightforward proposition. Is it possible for you to build a model that can differentiate between “Real “news and “Fake” news? A proposed work on assembling a dataset of both fake and real news and employing a Naive Bayes classifier in order to create a model to classify an article into fake or real based on its words and phrases.

Our project aims to use Natural Language Processing and various machine learning algorithms using sci-kit libraries from python to detect fake news directly, based on the text content a user enters.

Keywords:

Machine learning,python,Natural language processing,algorithms,Fake News

REQUIRED TOOLS

- PYTHON 3.X
- NLTK
- NUMPY
- SKLEARN
- SCIPY
- PANDAS
- SPACY

FEASIBILITY ANALYSIS

The idea doesn't seem to be very far-fetched, rather a very realistic approach in tackling one of the most common stigmas of society that is fake news.

Though the idea is simple it still requires some time to fully develop this project.

The aim is to create an app which would be compatible on every smartphone device or a website with the same level of compatibility and even on browsers.

The application can be made with the help of React and is definitely viable and the website can be made with the help of HTML and related web development tools which is again not that hard to access.

In a nutshell the resources to make this project are very much available which makes it a very feasible project. The app and website both would be free to access which would provide a vast exposure to the masses.

SCOPE/OBJECTIVE

The project aims at making use of various machine learning algorithms as well as making use of python developer tools. The project is not a very far-fetched idea, the need for such projects is growing exponentially in the market as the number of fake news sites are growing at a rapid rate.

The development of project goes through the following stages:

- **Collection of Data:**
  In this stage the data is collected as to train our machine learning algo for correctly differentiating between a real and a fake news. This source of data is easily accessible and can be found on any open source platform like Kaggle.
- **Training The Data Set:**
  The data set will be trained with the help of Aggressive-Passive machine learning algorithm. This algorithm is efficient for projects like these because this algorithm tackles a very large amount of data which would be used in the project.
- **Data Purification:**

The number of total tokens considered will be considerably reduced by eliminating stop words (such as “as” “is” “the” etc.. that are not believed to be causal on the article classification. To further purify the data, the extracted data will be manually purged of any unreadable or non-English characters or boiler-plate article headings.

Therefore by making use of simple machine learning algorithms and natural language processing techniques the project will be developed not taking much of the time. The project will be highly effective and functional and will be able to handle the large amount of data as well.

ANALYSIS

The aim of the project is to identify fake and real news efficiently by using Natural Language Processing Techniques and some machine learning algorithm

DEFINITION OF PROJECT

The project is mainly back end oriented and requires very less work on the front end. Since various Python libraries and machine learning algorithms will be used in this project therefore it is so dependent on the back end.

METHODOLOGY:

The methodology that will be followed in this project is quite simple.

- First, the collection of data will be done from some open source platform such as kaggle.

- Then comes training these data sets based on the machine learning algorithm or classifiers like passive aggressive classifier and naive bayes classifier.

- After training the data, the next step will be testing the project. The required outcome is that the model should be able to identify between real and fake news easily at the speed of a click.

- After the testing phase, the project will start taking shape in terms of user interface so that the users would be easily able to use the functions of the project without much of a hassle.
- The user interface will not be intricate, rather very decent and simple. The tools which would be used to design the interface would be HTML and CSS.

The timing of the analysis will be quite short as we are going to analyse the model on some predefined data sets.

DESIGN

DEFINITION:

The main design of the project will be focused on the front end as the project is mainly back end oriented.

The design will be done on the user interface of the project. The idea is to keep the design of the use interface as simple as possible because the user should be able to use the application or website without much of the trouble.

TOOLS:

The tools that will be used to develop the user interface are Flask,HTMl and CSS3.

LAYOUT OF DESIGN:

The user interface will contain an input box that will act as an area where users would input the news which they want to know about is fake or real. Then there will be a button that will act as the trigger to start the process of detection. Then beside that button will be a reset button that would help users to clean the input data which was entered earlier so that they can enter new data.

IMPLEMENTATION AND TESTING

IMPLEMENTATION:

- PreProcessing The Text:

  The performance of the test classification model is highly dependent on the words in a corpus and the features created from those words.Common words(otherwise known as stopwords) and other "noisy" elements increase feature dimensionality but do not usually help to differentiate between documents.

- Feature Extraction:

  To analyze and model text after it has been preprocessed, it must first be converted into features.Techniques include Bag of Words and TfiDfVectorizer.

  Bag of Words:This model analyzes the text from all input documents and converts it in a Bag-of-Words form.For example, for more than one text, we can have one bag of words which will contain all distinct words from all texts in one bag.

  Term Frequency-Inverse Document Frequency(TF-IDF): It increases the proportionality with the number of items a word appears in a document, but is offset by its frequency in the overall corpus. While TF-IDF is a good basic metric for extracting descriptive terms, it does not take into consideration a words' position or context.

- **Classification:**

  Two classifiers are being used in this project to train and test data and those are Naive Bayes Classifier and Passive Aggressive Classifier.

  Naive Bayes Classifier: In machine learning, Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with powerful(naive) independent assumptions between the features.

Passive Aggressive Classifier:This algorithm remains passive for a correct classification outcome, and turns aggressive in the event of a miscalculation.Its purpose is to make updates that correct the loss,causing very little change in the norm of the weight vector.

TESTING

The testing of the project will be done against the classifier that we are using in our model which are Naive Bayes and Passive Aggressive. Not just testing but the training of the model will also be done on the basis of these two classifiers.

Various data sets will be fed to the model and the expected outcome would be “real” or “fake”.

LIMITATIONS

The limitation of this project so far is that it is functional only for some predefined data sets against which the model would be trained. The project will be dynamic to a less extent.

FUTURE SCOPE OF THE PROJECT

The future scope of the project is to make it as dynamic as possible with minimum to zero possibilities of discrepancies. After that the project would be able to identify any type of data as “real” or “fake” from across the world in just a click away.
The application or the website would be free to access and would be accessible from any part of the world.

CONCLUSIONS

REFERENCES

1. Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc. Ng, Boneh, (2017), CS 229: Machine Learning. Course Material. .
2. Zhou, X., Jain, A., Phoha, V.V., Zafarani, R.: Fake news early detection: a theory-driven model. arXiv preprint arXiv:1904.11679 (2019)
3. Zhou, X., Zafarani, R.: Fake news: a survey of research, detection methods, and opportunities. arXiv preprint arXiv:1812.00315 (2018)
4. Zhou, X., Zafarani, R., Shu, K., Liu, H.: Fake news: Fundamental theories, detection strategies and challenges. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, pp. 836–837. ACM (2019)
5. Bovet, A., Makse, H.A.: Influence of fake news in Twitter during the 2016 us presidential election. Nat. Commun. 10(1), 7 (2019)