Project

on

FAKE NEWS DETECTION

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

Bachelor of Technology in Computer Science & Engg.

Under The Supervision of
Mr. P. Raja Kumar
Associate Professor

Submitted By

Shubham Bist
18SCSE1010444

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA
DECEMBER, 2021

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

I/We hereby certify that the work which is being presented in the thesis/project/dissertation, entitled **"FAKE NEWS DETECTION"** in partial fulfillment of the requirements for the award of the Bachelor of Technology submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of month, Year to Month and Year, under the supervision of Name… Designation, Department of Computer Science and Engineering/Computer Application and Information and Science, of School of Computing Science and Engineering , Galgotias University, Greater Noida

The matter presented in the thesis/project/dissertation has not been submitted by me/us for the award of any other degree of this or any other places.

Shubham Bist
18SCSE1010444

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Mr. P. Raja Kumar
Associate Professor

CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of Shubham Bist(18SCSE1010444) has been held on _____ and his/her work is recommended for the award of Bachelor Of Technology.

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date: November, 2013
Place: Greater Noida

Abstract

Fake news has quickly become a society problem, being used to propagated false or rumour information in order to change people's behaviour, this topic on social media has recently attracted tremendous attention.

We are using the Naïve bays and passive aggressive classifier algorithm which can predict with an accuracy of roughly around 86%. This could help the novice project creator and could assist them in the planning for their crowd funding project. The future work in the project is that we might need to tune the model again as news in something which we couldn't predict.

Contents

List of Table

List of Figures

Acronyms

ML	MACHINE LEARNING
TF-IDF	TEAM FREQUENCY- INVERSE DOCUMENT FREQUENCY
NLP	NEURAL LANGUAGE PROCESSING
HTML	HYPER TEXT MARKUP LANGUAGE
POS	POINT OF SALE

CHAPTER-1 Introduction

So first let’s see what really is fake news? Stories that look like real news but are hoax, disinformation. It typically appears on websites that may look professional. Moreover these topics are related to what is trending on social media already. These stories have outrageous headlines designed to get people to click them and believe them.

Sometimes these stories are created to attract an audience to particular website and gain profit.
According to the current situation of web-based media and each sort of web related thing, individuals are completely relied upon that occasionally we don’t realize that each news and article are not a genuine article which occurred in the whole world however we were had confidence in that online media is the biggest client base stage which comprises the news that occasionally genuine or counterfeit this framework distinguishes that each sort of phony and genuine news with an incredible foundation of information science and furthermore utilizes a lot of dataset which comprises bunches of information related information by logical stages. The basic countermeasure of comparing websites against a list of labeled fake news sources is inflexible, and so a machine learning approach comes into prime.

CHAPTER-2 Literature Survey

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 thinks of the uses of NLP (Natural Language Processing) methods for identifying the 'phony news', that is, deceiving reports that come from non-trustworthy sources Combatting the fake news is a classic text classification project with a straightforward proposition. It is workable for you to assemble a model that can separate between "Genuine "news and "Phony" news? A proposed work on gathering a dataset of both phony and genuine news and utilizing a Naive Bayes classifier to make a model to order an article into phony or genuine dependent on its words and expressions.

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 literature review

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.

Now talking about the early research of this project: Conroy, Rubin, and Chen diagram a few methodologies that appear to be encouraging towards the point of consummately ordering the deceptive articles. They note that straightforward substance related n-grams and shallow grammatical forms (POS) labeling have demonstrated deficiency for the characterization task, regularly neglecting to represent significant setting data. Or maybe, these techniques have been demonstrated helpful just coupled with more mind boggling strategies for examination.
Profound Syntax investigation utilizing Probabilistic Context Free Grammars (PCFG) have been demonstrated to be especially important in blend with n-gram strategies. Feng, Banerjee, and Choi can accomplish 85%-91% precision in misdirection related characterization undertakings utilizing on the web audit corporation.

Feasibility Analysis

There are two important things to do before one sets out to pursue any particular task. The first is to decide if the specific undertaking can be effectively done; the second is to decide if the specific errand merits doing. Planning for this task achievability examination is significant. A feasibility analysis helps to evaluate the success of a process and therefore perceived objectivity is an essential factor for potential investors and lending institutions. By running various examinations and quality checks on our project, we reached on the conclusion that our project is feasible in all these five areas - Technical Feasibility, Economic Feasibility , Legal Feasibility Operational Feasibility and Scheduling Feasibility

The Technical Feasibility assessment mainly focuses on the technical resources available to the organisation. It helps the organizations to find out whether the technical resources meet the capacity and whether or not the team is capable of converting the ideas into working systems.

This Economic Feasibility evaluation includes a cost/'advantage examination of the undertaking and assisting the association with deciding the viability. Cost and advantages related with a venture before monetary assets are dispensed.

Proposed Methodology

Preprocessing the text:

The presentation of the test characterization model is profoundly reliant on the words in a corpus and the highlights made from those words.Common words(otherwise known as stopwords) and other "loud" components increment include dimensionality however don't normally assist with separating between reports.

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 decent fundamental measurement for extricating illustrative terms, it doesn't contemplate a words' position or settings.

Dataset

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.

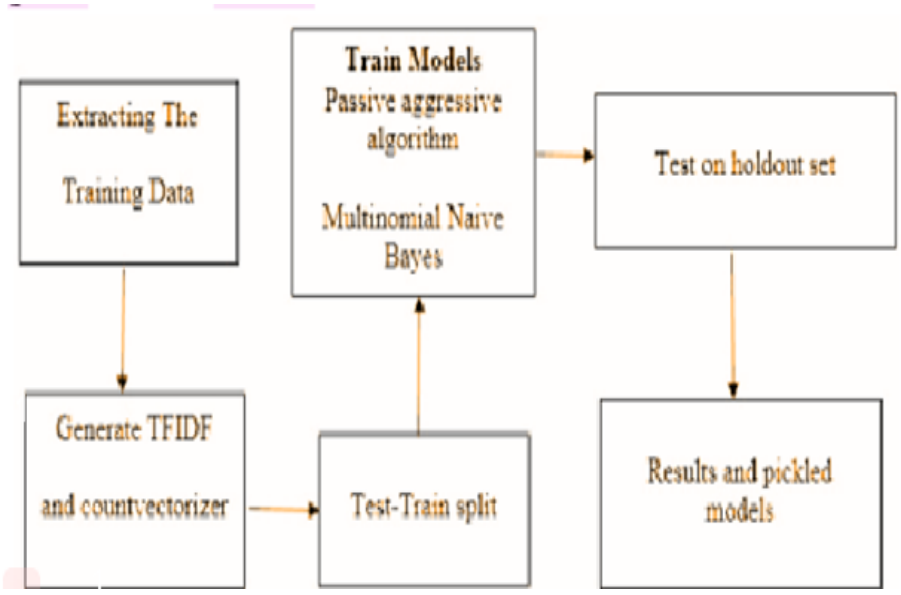
Data Purification:

The total number of tokens will be substantially reduced by collapsing stop words (such as "as" "is" "the" etc.. that are not believed to be causal on the article classification. To additionally filter the information, the extricated information will be physically cleansed of any mixed up or non-English characters or standard 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.

Proposed Model

This model is built based on the countvectorizer or a tfidf matrix (i.e) word tallies relative to how often they are used in other articles in your dataset) can help. Since this issue is a sort of text characterization, Implementing a Naive Bayes classifier will be best as this is standard for text-based preparing. The actual goal lies in developing a model which would be the text transformation (countvectorizer vs tfidf vectorizer) and henceforth picking up which type of text to be made use of. (headlines vs full text). Presently the subsequent stage is to extricate the most ideal highlights for countvectorizer or tfidf-vectorizer, this is finished by utilizing a n-number of the most utilized words, as well as expressions, lower packaging or not, predominantly eliminating the stop words which are normal words, for example, "the", "when", and "there" and just utilizing those words that show up at any rate a given number of times in a given content dataset. We first need to filter the dataset as it contains a lot of redundancy. we will be keeping only those project data whose goal status is successful or unsuccessful or fake or real. we will be removing all those projects whose status is canceled, live, undefined, or suspended. Also, we will be removing those projects whose goal is completed after the deadline.



Algorithms

The labels are in equal amounts, that is the number of true and false labels are in the same proportion.

Confusion matrix

A greatly improved approach to assess the presentation of a classifier is to take a gander at the disarray network. The overall thought is to tally the occasions examples of class A are named class B. For instance, to know the occasions the classifier befuddled pictures of 5s with 3s, you would glance in the fifth line and third section of the disarray network. 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 Al, Naive Bayes classifiers are a group of straightforward "probabilistic classifiers" in light of applying Bayes' hypothesis with powerful(naive) autonomous suppositions between the highlights. The formula used by this algorithm is as below:

P(A/B)=

P(BA) P(A)

P(B)

Passive Aggressive Classifier:

This calculation stays uninvolved for a right. grouping result, and turns forceful in case of a miscalculation. Its object is to make refr shes that right the loss causing almost no adjustment in the standard of the weight vector.

Passive: If the forecast is right, keep the model and don't roll out any improvements. i.c., the information in the model isn't sufficient to bring

about any adjustments in the model. Aggressive: If the expectation is erroneous, make changes to the model. i.e., some change to the model may address it.

Let's suppose to have a dataset:

(X = {x。 , × ₁, ... , *,, ... } where x¡ € Rª Y = {yo, y₁,... Y... } where y, E {-1, +

The list has been decided to stamp the worldly

measurement. For this situation, truth be told, the examples can keep showing up for an uncertain time. Obviously, in the event that they are drawn from a similar information producing dissemination, the calculation will continue learning (likely without huge boundary alterations), yet in the event that they are drawn from a totally unique appropriation, the loads will gradually fail to remember the past one and gain proficiency with the new conveyance. For straightforwardness, we likewise accept that we're working with a twofold order dependent on bipolar marks.

Given a weight vector w, the estimation is essentially gotten as:

$$\tilde{y}t = \text{sign}(w.xt)$$

Every one of these calculations depend on the Hinge misfortune work (a similar utilized by SVM):

$$L(A) = \max(0,1-y-f(xt;))$$

The estimation of L is limited between 0 (which means wonderful match) and K relying upon f(x(t),0) with K>0 (totally wrong expectation). A Passive-Aggressive calculation works conventionally with this update rule: can keep showing up for an uncertain time. Obviously, in the event that they are drawn from a similar information

producing dissemination, the calculation will continue learning (likely without huge boundary alterations), yet in the event that they are drawn from a totally unique appropriation, the loads will gradually fail to remember the past one and gain proficiency with the new conveyance. For straightforwardness, we likewise accept that we're working with a twofold order dependent on bipolar marks.

Given a weight vector w, the estimation is essentially gotten as:

$$\tilde{y}t = \text{sign}\ (w.xt)$$

Every one of these calculations depend on the Hinge misfortune work (a similar utilized by SVM):

$$L(A) = \max(0,1\text{-}y\text{-}f(xt;))$$

The estimation of L is limited between 0 (which means wonderful match) and K relying upon f(x(t),0) with K>0 (totally wrong expectation). A Passive-Aggressive calculation works conventionally with this update rule.

Methods and Software

The dataset which we have from Kaggle consists of records of all the projects launched on this website from 2010 to 2018. The CSV mat of the dataset consists of 378661 rows and 15 columns(features). However, there are many features in the dataset which tells us about the active stage of a project like so we have to get rid of those features because we want to predict the state of the project before it's release, so we won't be known about these features at the time of launching hence we can't use these features for training purpose. In the next phase of data cleaning, there are two features which tell us about the launching date and deadline of a project, they are not in the proper format so we changed those in DateTime format to make operations easier on those columns. Also in the launched column, time is given with date and the time of launching a project seems of no use to us so we will also eliminate that.

The goal of our problem is to predict whether the project will be successful or fail i.c., whether it will be able to collect the required money or not? There are many projects in our dataset which are still inactive, so we will eliminate those instances too as we can't use them to train our model because their outcome is not known yet. We will only use those instances to train our model which are either successful or failed as we want to predict these two states only. Packages Used: Numpy, Pandas, Scipy

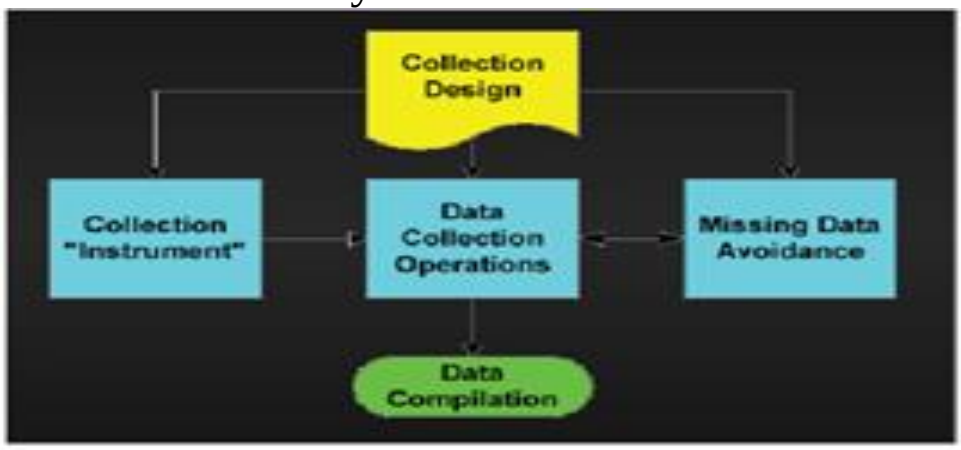
Feature Engineering is the most critical part of any machine learning project. We get what we feed to our model. If we want to get better accuracy then we have to figure out the most important features for our project and will train only using those features. For this problem, we had generated a new feature which tells us how many days a particular project is active i.c., what is the gap between the launch date and deadline. There are two categorical features in our final set of features, so we have to encode them too as we can't give categorical values for training our model. Here, the values for both the features are nominal i.e., there isn't any order associated with them hence label encoding would be most appropriate for this type of categorical features.

Feature scaling is the most important step for any machine learning project, there are several algorithms which perform much better when our features are on the same scale. We have used Standard scaling for this project, all of our features follow a standard normal distribution.

Packages Used: Numpy, Pandas, Sci-kit Learn Now as we have prepared our dataset for training purposes and all set to go. Next, we will split our data into training and testing data and will perform a quick check on 4 major classification algorithms. We will note down the scores then we will move further to perform hyper-parameter tuning to find the best parameters and cross-validation score for all the algorithms and will finalize the one giving the maximum cross-validation score.

Experimental Analysis

We have applied several results to our dataset and the results we got are shown in Fig 2, this is not cross-validation or score but a score on a random portion of data using default parameters which gives us an overview of which model could perform better on this type of dataset. The final score may increase or decrease after applying cross-validation. Next, we will perform hyperparameter tuning for all 3 models and will choose the best parameters and best model as the final model which gives us maximum accuracy.



Future Work

The result we have achieved is encouraging and we will try to improve our model. It could be increased by testing various powerful boosting algorithms like voting classifier and boost. We will try to speed up our model using boosting algorithms like AdaBoost. Also. In the future, if the trend changes then the model also needs to be tuned again. If not trained the model won't be able to capture that trend and might result in the wrong output. There is a high chance that the project will be of great success as crowdfunding is on of the best approach to showcase the ideas and generate funds for the project as we can see that only 35% of the project can reach their goal amount so this project will be of great help to help to predict the success of this project.

Conclusion

The result we have achieved is encouraging and we will try to improve our model. It could be increased by testing various powerful boosting algorithms like voting classifier and boost. We will try to speed up our model using boosting algorithms like AdaBoost. Also, in future ,if the trend changes then the model also needs to be tuned again. If not trained the model won't be able to capture that trend and might result in the wrong output. There is a high chance that the project will be of great success as crowdfunding is on of the best approach to showcase the ideas and generate funds for the project as we can see that only 35% of the project can reach their goal amount so this project will be of great help to help to predict the success of this project.

References

[1] dataset(online) Available at-https://www.kaggle.com/kemical/kickstarter-projects

[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 Conferenece 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)

[6] 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.

[7]https://towardsdatascience.com/full-pipeline-project-python-ai-for-detecting-fake-news-with-nlp-bbb 1 eec4936d

[8]https://medium.com/analytics-vidhya/fake-news-detection-using-python-learn-data-science-in-2020-80a45727a67b

[9] https://www.sciencedirect.com/