**INTRODUCTION:**

Fake news is a piece of information that consists of planned misinformation, and it is circulated throughsome medium, especially social media. The unverifiedinformation is spread through traditional news platforms, in the pursuit of some agendas, concerted efforts are designed to draft something out of nothing. Today in the era of tech revolution Things are moving in a different direction, fighting an utterly different war, and on an entirely different level. This fake news is controlled by various masses and business tycoons to fulfil political agenda and they are often perceived as one of the greatest threats to democracy, independence and national interests.

**FAKE NEWS DETECTION IN SOCIAL MEDIA:**

Fake news and hoaxes have been there since before the advent of the Internet. Thewidely accepted definition of Internet fake news is: fictitious articles deliberately fabricated to deceive reader’s. Social media and news outlets publish fake news toincrease readership or as part of psychological warfare. In general, the goal is profiting through click baits. Clickbaits lure users and entice curiosity with flashyheadlines or designs to click links to increase advertisements revenues. This exposition analyses the prevalence of fake news in light of the advances incommunication made possible by the emergence of social networking sites. Thepurpose of the work is to come up with a solution that can be utilized by users to detect and filter out sites containing false and misleading information. We use simpleand carefully selected features of the title and post to accurately identify fake posts.The experimental results show a 99.4% accuracy using logistic classifier.Automatic Online Fake News Detection Combining Content and Social Signal.

The proliferation and rapid diffusion of fake news on the Internet highlight the need ofautomatic hoax detection systems. In the context of social networks, machinelearning (ML) methods can be used for this purpose. Fake news detection strategiesare traditionally either based on content analysis (i.e., analysing the content of thenews) or - more recently - on social context models, such as mapping the news‟diffusion pattern. In this paper, we first propose a novel ML fake news detectionmethod which, by combining news content and social context features, outperforms existing methods in the literature, increasing their already high accuracy by up to4.8%. Second, we implement our method within a Facebook Messenger chatbot andvalidate it with a real-world application, obtaining a fake news detection accuracy of81.7%.

In recent years, the reliability of information on the Internet has emerged as a crucialissue of modern society. Social network sites (SNSs) have revolutionized the way inwhich information is spread by allowing users to freely share content. As aconsequence, SNSs are also increasingly used as vectors for the diffusion ofmisinformation and hoaxes. The amount of disseminated information and the rapidityof its diffusion make it practically impossible to assess reliability in a timely manner,highlighting the need for automatic hoax detection systems.As a contribution towardsthis objective, we show that Facebook posts can be classified with high accuracy ashoaxes or non-hoaxes on the basis of the users who "liked" them. We present twoclassification techniques, one based on logistic regression, the other on a noveladaptation of Boolean crowdsourcing algorithms. On a dataset consisting of 15,500Facebook posts and 909,236 users, we obtain classification accuracies exceeding99% even when the training set contains less than 1% of the posts. We further showthat our techniques are robust: they work even when we restrict our attention to theusers who like both hoax and non-hoax posts. These results suggest that mappingthe diffusion pattern of information can be a useful component of automatic hoax Detection systems.

**METHODOLOGY**

**EXISTING SYSTEM**

There exists a large body of research on the topic of machine learning methods fordeception detection, most of it has been focusing on classifying online reviews andpublicly available social media posts. Particularly since late 2016 during theAmerican Presidential election, the question of determining 'fake news' has alsobeen the subject of particular attention within the literature. Conroy, Rubin, andChen outline several approaches that seem promising towards the aim of perfectlyclassify the misleading articles. They note that simple content-related n-grams andshallow parts-of-speech tagging have proven insufficient for the classification task, often failing to account for important context information. Rather, these methodshave been shown useful only in tandem with more complex methods of analysis.Deep Syntax analysis using Probabilistic Context Free Grammars have been shownto be particularly valuable in combination with n-gram methods. Feng, Banerjee, andChoi are able to achieve 85%-91% accuracy in deception related classification tasksusing online review corpora.

**PROPOSED SYSTEM**

In this paper a model is build based on the count vectorizer or a tfidf matrix ( i.e )word tallies relatives to how often they are used in other articles in your dataset ) canhelp . Since this problemis a kind of text

classification, implementing a Naive Bayesclassifier will be best as this is standard for text-based processing. The actual goalis in developing a model which was the text transformation (count vectorizer vs tfidf vectorizer) and choosing which type of text to use (headlines vs full text). Now thenext step is to extract the most optimal features for count vectorizer or tfidf-vectorizer,this is done by using a n-number of the most used words, and/or phrases, lowercasing 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 givennumber of times in a given text dataset.

**SYSTEM REQUIREMENTS**

**HARDWARE REQUIREMENTS:**

* System – Pentium-IV
* Speed – 2.4GHZ
* Hard disk – 40GB
* Monitor – 15VGA color
* RAM – 512MB

**SOFTWARE REQUIREMENTS:**

* Operating System – Windows XP
* Coding language – PYTHON

**SOFTWARE ENVIRONMENT**

PYTHON

Python is a high-level, interpreted, interactive and object-oriented scripting language.Python is designed to be highly readable. It uses English keywords frequently whereas other languages use punctuation, and it has fewer syntactical constructions than otherlanguages. Python is Interpreted – Python is processed at runtime by the interpreter.You do not need to compile your program before executing it. This is similar toPERL and PHP.

* Python is Interactive – You can actually sit at a Python prompt and interactwith the interpreter directly to write your programs.
* Python is Object-Oriented – Python supports Object-Oriented style orTechnique of programming that encapsulates code within objects.
* Python is a Beginner’s Language – Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

History of Python

Python was developed by Guido van Rossum in the late eighties and early nineties at the National Research Institute for Mathematics and Computer Science in theNetherlands.Python is derived from many other languages, including ABC, Modula-3, C, C++,Algol-68, small talk, and Unix shell and other scripting languages.Python is copyrighted. Like Perl, Python source code is now available under theGNU General Public License (GPL).Python is now maintained by a core development team at the institute, althoughGuido van Rossum still holds a vital role in directing its progress.

PYTHON FEATURES

Python’s features include –

* Easy-to-learn – Python has few keywords, simple structure, and a clearly
* Defined syntax. This allows the student to pick up the language quickly.
* Easy-to-read – Python code is more clearly defined and visible to the eyes.
* Databases − Python provides interfaces to all major commercial databases.
* GUI Programming − Python supports GUI applications that can be createand ported to many system calls, libraries and windows systems, such as
* Windows MFC, Macintosh, and the X Window system of Unix.
* Scalable − Python provides a better structure and support for large program than shell scripting.

**MODULES**

**A.** Data Use

**B.** Preprocessing

**C.** Feature Extraction

**D.** Training the Classifier

**MODULES DESCRIPTION**

**A. Data Use**

So, in this project we are using different packages and to load and read the data set we are using pandas. By using pandas, we can read the .csv file and then we can display the shape of the dataset with that we can also display the dataset in the correct form. We will be training and testing the data, when we use supervised learning, it means we are labelling the data. By getting the testing and training data and labels we can perform different machine learning algorithms but before performing the predictions and accuracies, the data is need to be preprocessing i.e., the null values which are not readable are required to be removed from the data set and the data is required to be converted into vectors by normalizing and tokening the data so that it could be understood by the machine. Next step is by using this data, getting the visual reports, which we will get by using the Mat Plot Library of Python and Sickit Learn. This library helps us in getting the results in the form of histograms, pie charts or bar charts.

**B.Preprocessing**

The data set used is split into a training set and a testing set containing in Dataset 1 -3256 training data and 814 testing data and in Dataset II- 1882 training data and 471 testing data respectively. Cleaning the data is always the first step. In this, those words are removed from the dataset. That helps in mining the useful information. Whenever we collect data online, it sometimes contains the undesirable characters like stop words, digits etc. which creates hindrance while spam detection. It helps in removing the texts which are language independent entities and integrate the logic which can improve the accuracy of the identification task.

**C. Feature Extraction**

Feature extraction s the process of selecting a subset of relevant features for use inmodelconstruction. Feature extraction methods helps in to create an accuratepredictive model. They help in selecting features that will give better accuracy. Whenthe input data to an algorithm is too large to be handled and it’s supposed to beredundant then the input data will be transformed into a reduced illustration set offeatures also named feature vectors. Altering the input data to perform the desiredtask using this reduced representation instead of the full-size input. Featureextraction is performed on raw data prior to applying any machine learning algorithm,on the transformed data in feature space.

**D. Training the Classifier**

As In this project I am using Scikit-Learn Machine learning library for implementingthe architecture.Scikit Learn is an open-source python Machine Learning librarywhich comes bundled in 3rd distribution anaconda. This just needs importing thepackages and you can compile the command as soon as you write it. If thecommand doesn’t run, we can get the error at the same time. I am using 4 differentalgorithms and I have trained these 4 models i.e., Naïve Bayes, Support VectorMachine, K Nearest Neighbours and Logistic Regression wic are very popularmethods for document classification problem. Once the classifiers are trained, wecan c heck the performance of the models on test-set. We can extract the wordcount vector for each mail in test-set and predict it class with the trained models.

**Algorithms**

**Naive Bayes**

* One of supervised learning algorithm based on probabilistic classification technique.
* It is a powerful and fast algorithm for predictive modelling.
* In this project, I have used the Multinomial Naive Bayes Classifier.

**Support Vector Machine- SVM**

* SVM’s are a set of supervised learning methods used for classification, and regression.
* Effective in high dimensional spaces.
* Uses a subset of training points in the support vector, so it is also memory efficient.

**Logistic Regression**

* Linear model for classification rather than regression.
* The expected values of the response variable are modelled based on combination of values taken by the predictors.

**RESULTS AND DISCUSSION**

* Algorithm’s accuracy depends on the type and size of your dataset. More the data, more chances of getting correct accuracy.
* Machine learning depends on the variations and relations
* Understanding what is predictable is as important as trying to predict it.
* While making algorithm choice , speed should be a consideration factor.

**REQUIREMENT ANALYSIS**

Requirement analysis, also called requirement engineering, is the process ofdetermining user expectations for a new modified product. It encompasses the tasksthat determine the need

for analysing, documenting, validating and managingsoftware or system requirements. The requirements should be documentable,actionable, measurable, testable and traceable related to identified business needsor opportunities and define to a level of detail, sufficient for system design.

**FUNCTIONAL REQUIREMENTS**

It is a technical specification requirement for the software products. It is thefirst step in the requirement analysis process which lists the requirements ofparticular software systems including functional, performance and securityrequirements. The function of the system depends mainly on the quality hardwareused to run the software with given functionality

**Usability**

It specifies how easy the system must be use. It is easy to ask queries in anyformat which is short or long, porter stemming algorithm stimulates the desiredresponse for user.

**Robustness**

It refers to a program that performs well not only under ordinary conditions butalso under unusual conditions. It is the ability of the user to cope with errors forirrelevant queries during execution.

**Security**

The state of providing protected access to resource is security. The systemprovides good security and unauthorized users cannot access the system there by providing high security.

**Reliability**

It is the probability of how often the software fails. The measurement is oftenexpressed in MTBF (Mean Time Between Failures). The requirement is needed inorder to ensure that the processes work correctly and completely without beingaborted. It can handle any load and survive and survive and even capable of workingaround any failure.

**Compatibility**

It is supported by version above all web browsers. Using any web servers likelocalhost makes the system real-time experience.

**Flexibility**

The flexibility of the project is provided in such a way that is has the ability to run ondifferent environments being executed by different users.

**Safety**

Safety is a measure taken to prevent trouble. Every query is processed in a securedmanner without letting others to know one’s personal information.

**NON- FUNCTIONAL REQUIREMENTS**

**Portability**

It is the usability of the same software in different environments. The projectCan be run in any operating system.

**Performance**

These requirements determine the resources required, time interval,throughput and everything that deals with the performance of the system.

**Accuracy**

The result of the requesting query is very accurate and high speed of retrieving information. The degree of security provided by the system is high and effective.

**Maintainability**

Project is simple as further updates can be easily done without affecting its

stability. Maintainability basically defines that how easy it is to maintain the system. Itmeans that how easy it is to maintain the system, analyse, change and test theapplication. Maintainability of this project is simple as further updates can be easily **d**one without affecting its stability.

**SYSTEM DESIGN AND TESTING PLAN**

**INPUT DESIGN**

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system.The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

* What data should be given as input?
* How the data should be arranged or coded?
* The dialog to guide the operating personnel in providing input.
* Methods for preparing input validations and steps to follow when errorOccur.

**OUTPUT DESIGN**

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users

and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user.Efficient and intelligent output design improves the system’s relationship to help user decision-making.The output form of an information system should accomplish one or more of the following objectives.

* Convey information about past activities, current status or projections of the Future.
* Signal important events, opportunities, problems, or warnings.
* Trigger an action.
* Confirm an action.

**SOCIAL FEASIBILITY**

The aspect of study is to check the level of acceptance of the system by the user.This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

**DATA FLOW DIAGRAM**

* The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
* The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
* DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.
* DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.
* It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration.