Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Engineering
School of Computer Science and Statistics

How can information literacy be incorporated and built into an environment that will provide consumers with a platform that will help segregate information and disinformation?

Yash Pandey
[redacted]

Supervisor- Dr Owen Conlan

30/04/2020

A dissertation submitted in partial fulfilment
of the requirements for the degree of
M.A.I. (Computer Engineering)

Student's Declaration

I hereby declare that this project is entirely my own work except for quotations and summaries which have been duly acknowledged. It has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at http://www.tcd.ie/calendar.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write.

Signature: Yash
Name of the Supervisor: Dr Owen Conlan
Date:30th April 2020

Acknowledgement :

I would like to express my thanks and gratitude for the valuable contribution to my supervisor, Dr Owen Conlan, which helped in the development of this project from the initial stage till the end. Not only he has been very patient and guided me throughout during the project but also mentored me and gave feedback which greatly improved my work. Without his continuous support, advise and assistance, this project would not have been not possible.

I would also like to thank Dr Bilal Yousuf for his continuous guidance and valuable feedback which helped me in designing and implementing this project.

I acknowledge my sincere indebtedness and gratitude to my parents. I am really thankful for their sacrifice, patience, and understanding without which this would not have been possible. It's their devotion, support and faith in my ability which helped me achieve my dreams.

Lastly, I would like to thanks any person which contributes to my final year project directly on indirectly. I would like to acknowledge their comments and suggestions, which was crucial for the successful completion of this project.

Abstract :

People who are information literate are good at knowing when they have a need for information, identifying information needed to address a given problem or issue, finding needed information, evaluating the information, organizing the information, and using the information effectively. The major issue is the decreasing ability of a consumer to evaluate information and use it wisely, and this is mainly due to the rapid growth in the access of information thus the number of educated information consumers are decreasing.

This dissertation proposes a framework which will incorporate information literacy and build an environment that will provide consumers with a platform that will help segregate information and disinformation.

The first phase of the project details related state of the artwork that is ongoing or is already implemented int the field of information literacy, and focuses in the importance of information literacy and why is it needed.

The second phase of the project details the design of a Machine learning model and it addresses all the task which are to be completed in order to develop a machine learning model.

The third phase and the most important details the implementation of the model which happens in several complex steps which include which included data gathering, data processing, data wrangling, selecting and training the classifier models.

The final phase of the project evaluates the performance on a test dataset obtained from Snopes. It compares the accuracy of the models implemented further discusses the key findings and limitations. In the end, conclusion and future prospects about the project are discussed.

Table of contents :

b) List of Figures(63)

1) INTRODUCTION

1.a) Motivation

1.b) Background

1.c) Aims and Objectives

1.d) Dissertation Overview

1.a) MOTIVATION

As the convergence of media is growing, the boundaries between information and media are becoming increasingly blurred and hence we can no longer regard new media as a matter of just technology or information. Hence a major issue arises, that is the decreasing ability of a consumer to evaluate information and use it wisely, and this is mainly due to the rapid growth in the access of information thus the number of educated information consumers are decreasing.

People who are information literate are good at "knowing when they have a need for information. identifying information needed to address a given problem or issue, finding needed information, evaluating the information, organizing the information, and using the information effectively to address the problem or issue at hand. Hence there is a requirement of a framework that will enhance the consumers' ability to evaluate information and disinformation and make them information literate.

1.b) BACKGROUND

The debate about digital technology and education has moved beyond the question of basic access. Attention is now focusing on the issue of what young people need to know about the technology that is, the forms of competence and understanding they need if they are going to use technology effectively and critically.

The main problems in evaluating information lie in the difficulty of assessing the credibility and originality of information and the professional integrity of its presentation. The boundary between traditional news that is information and user-generated content that is disinformation has also become increasingly blurred. Hence there is a need for the development of automated tools and

technologies to assist the consumers in identifying the information out of disinformation by verifying, checking and filtering the content.

The most tried and tested method to make someone understand something which in this case would be injecting information literacy in a very short time and an impactful way is though visualisations. The goal of visualization is to aid our understanding of data by leveraging the human visual system's highly tuned ability to see patterns, spot trends, and identify outliers. To create a visualisation number of factors are to be considered such as what questions are being answered by the visualisation, the type of data that is being used, what visual encoding should be used given a particular data which would again depend on various factors such as colour, size, shape and position.

1.c) Research Question and Objectives

The idea is very simple, that there are definitely certain distinctive differences between information and disinformation or we might as well say a piece of true news and a piece of fake news. But as the convergence of media is growing, the boundaries between information and disinformation are becoming increasingly blurred, which in turn is leading to a decreased ability of a consumer to evaluate information and use it wisely. People who are information literate are excellent at 'knowing where they need information. identifying the information needed to address the situation or issue, locating the information needed, assessing the details, managing the information, and utilizing the information efficiently to resolve the issue or issue at hand. Hence this brings us to the research question:

How can information literacy be incorporated and built into an environment that will provide consumers with a platform that will help segregate information and disinformation?

To address the research question, objectives need to be defined. The objectives of this project are:

1) To review the existing literature to explore, the boundaries between information and disinformation, difficulties that a consumer face to differentiate between information and disinformation, and different visualisations that can be used to fulfil our purpose.

2) To design a Machine learning model which will incorporate information literacy and build an environment that will provide consumers with a platform that will help segregate information and disinformation.

3) To implement the designed machine learning model through a pipeline of complex steps which included data gathering, data processing, data wrangling, selecting and training the classifier models.

4) To evaluate the performance of the models implemented and compare them.

1.4) DISSERTATION OVERVIEW

The goal of this study is to provide a thorough explanation and documentation of how information literacy can be implemented resulting in an environment which will enable customers to distinguish information and misinformation.

Chapter 2 offers an analysis of current literature and state-of-the-art, reflects on why knowledge literacy is required and explores certain state-of-the-art technology, based on the conclusions the experiment is designed and implemented

Chapter 3 gives an overview of the theory and design of the Model. It discusses the tasks needed to be addressed and the resources used for building the model.

Chapter 4 gives an overview of the implementation of the Model. It discusses the pipeline of complex steps involved while implementing the model.

Chapter 5 evaluates the performance of the models and compares them. It further discusses the key findings and limitations of the project.

Chapter 6 discusses the conclusion and future prospects of this project.

2) Literature review

a) Introduction

b) what is information Literacy?

c) Visualisation techniques

d) State of the Art

2) Literature Review

2.a) Introduction

Bruce a well-known Australian information literacy researcher, notes:
"The idea of information literacy, emerging with the advent of information technologies in the early 1970s, has grown, taken shape and strengthened to become recognized as the critical literacy for the twenty-first century. Sometimes interpreted as one of a number of literacies, information literacy is also described as the overarching literacy essential for twenty-first-century living. Today, information literacy is inextricably associated with information practices and critical thinking in the information and communication technology environment" (Bruce, 2002) [22]

In truth, since 1974, information literacy has become a field of growing importance for librarians and knowledge practitioners, and there is a considerable amount of literature on the topic. Although, the majority of publications have come from industrial, English-speaking nations, especially the United States and Australia.[26,27]

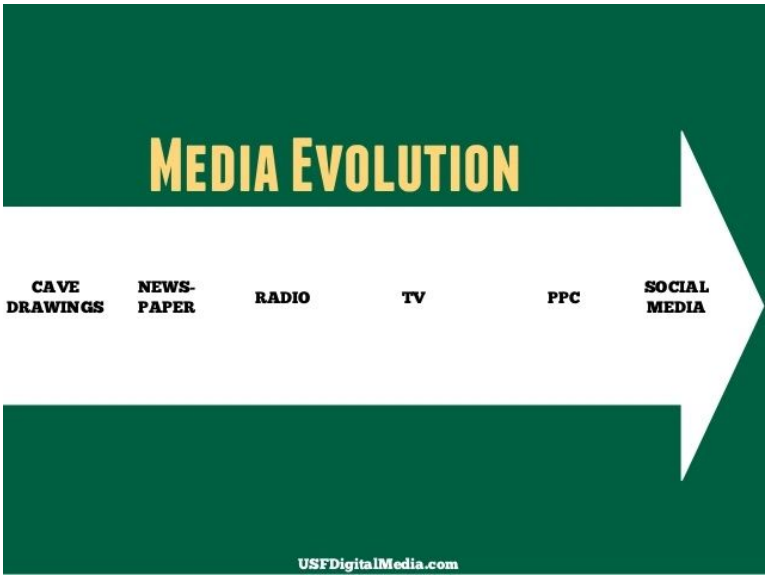

Fig1: Evolution of Media

The information literacy movement in the United States and Australia has been thoroughly studied and debated, and there have been major programs in other nations. The National Forum on Information Literacy was founded in 1989 in the United States, the Center for Information Literacy in 1998, while two sets of information literacy requirements were created for the school system and the higher education field. In December 2000, the United States Department of Education listed knowledge literacy in its National School Development Strategy as one of the five priorities. The value of students being able to view and analyze information is illustrated in a variety of other strategic materials. [31,32]

Examples of how information literacy programs and policies have been implemented in the United States can be seen at different levels. At the state level, for example, Colorado, Wisconsin and Oregon have implemented principles and many programs have been established by state-wide higher education networks, including the SUNY Knowledge Literacy Program, the California State University System Information Competence Project, Wisconsin and the University of Massachusetts. Standards have since been adopted for different colleges and universities. Some of these are Earlham College, Kings College, University of Louisville, University of Washington, University of Iowa, and Florida International University. [21]

The definition of information literacy has already permeated policy thought in Australia (Muir & Oppenheim, 2001) and has been illustrated in a range of important studies generated by the higher education sector and the government. The Council of Australian University Librarians (CAUL) has established information literacy principles adapted from those of the Association of College and Academic Libraries (ACRL) (CAUL, 2001) and has incorporated information literacy approaches into a variety of university institutional programs. For example, the Central Queensland University (CQU) distance education literacy program has been the focus of a number of grants and recognized as a flagship program both internationally and within Australia (Bruce & Candy, 2000), and the University of Ballarat's policy document identifies

information literacy as a key graduate outcome and as an integral part of the program.

There are also references to information literacy developments in Canada, China, Japan, Mexico, Namibia, New Zealand, Singapore and South Africa (Whitehead. & Quinlan, 2002; Spitzer, et al., 1998; Muir & Oppenheim, 2001; Rader, 2002a; Inoue, et al., 1997; Morgan, 2000; Moore, 2000; LIANZA, 2001; Hepworth, 2000a; Karelse, 2000). References to information literacy initiatives in Europe are, however, quite rare and fragmented. The majority of publications have come from the United Kingdom. Part of the problem of understanding European information literacy activities stems from the language barrier.[29,30]

For example, several information literacy programs have been recorded only in local languages-Danish, Dutch, Finnish, French, German, Norwegian, Spanish, Swedish and other languages but not in English. Examples among the most quoted and well-known programs are the EDUCATE and DEDICATE ventures funded by the European Union, led by Nancy Fjällbrant, the Chalmers University among Technology in Sweden, and the study proposal for a doctoral thesis on the quest for and usage of knowledge in the academic sense by Louise Limberg of the Swedish Institute of Library and Knowledge Studies. [28, 33]

In 2000 Charles Sturt University published Information literacy around the world: advances in programs and research, edited by Bruce and Candy, which includes the examples mentioned above. In addition, Mutch of Nottingham Trent University considers the nature of information literacy in the workplace (Bruce & Candy, 2000). However, during recent years there has been considerable interest in information literacy in Europe. This can be illustrated by the number of projects, conferences, workshops, working groups, an adaptation of information literacy competency standards, teaching initiatives in many institutions, development of Web sites and Web-based tutorials, and in the area of research.[36,37]

This chapter will give an overview of why information literacy is a necessity, and then pass on to various visualization strategies and how they can be used to

tackle the problem. The state-of-the-art developments that actually operate in these areas are being explored.

2.b) What is Information Literacy?

Information literacy is essential for today's learners, it encourages problem-solving strategies and analytical skills – asking questions and discovering responses, locating facts, developing ideas, analyzing sources and taking decisions that encourage active learners, productive participants, competent individuals and responsible citizens. [19]
It is at the core of the Curriculum for Consistency and Understanding through learning experiences and outcomes – the responsibility of all professionals.
They need to be able to recognize what is true and important not just for education, but also for studying, life and work.
Shigeru Aoyagi, Chief, Division of Basic Education, UNESCO, stated that:

*"**For all societies, Information Literacy is becoming an increasingly important component of not only literacy policies and strategies but also of global policies to promote human development**."*

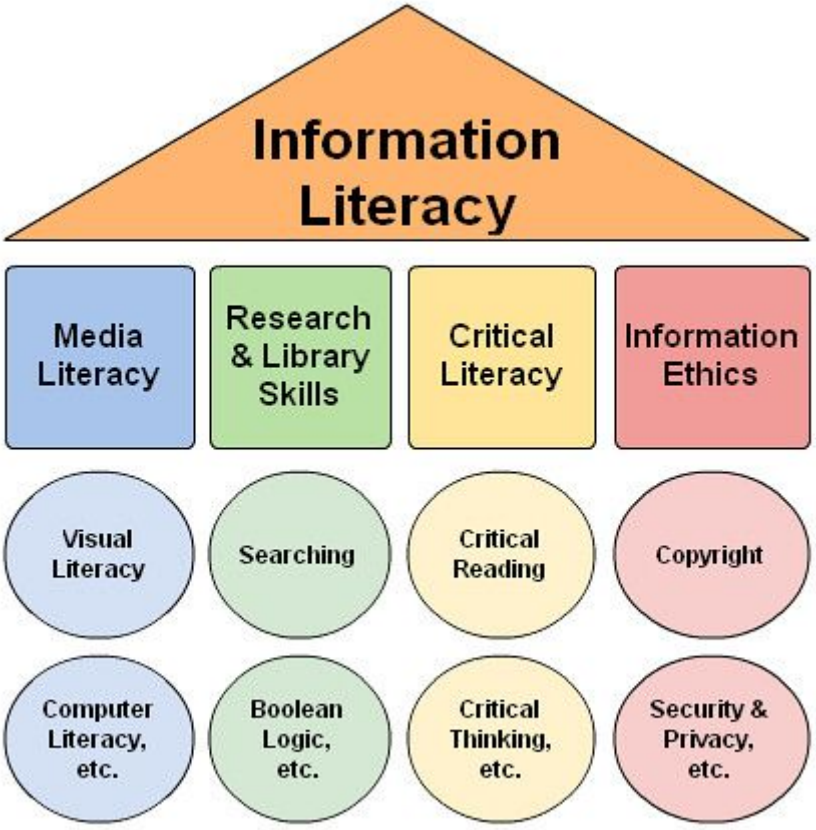

Fig2: information literacy pyramid

"Information literacy is knowing when and why you need information, where to find it, and how to evaluate, use and communicate it in an ethical manner."[18]

This definition implies several skills ... (or competencies) that are required to be information literate ... an understanding of:

- A need for information
- The resources available
- How to find information
- The need to evaluate results
- How to work with or exploit results
- Ethics and Responsibility for use
- How to communicate or share your findings
- How to manage your findings

Fig3: information literacy cycle

Information is all around us; it comes in different formats and is essential for our decision making and problem-solving activities whether at work, in education or at home. The difference is in why we need it, the level of information we need and how we are going to use it.

The debate about digital technology and education has moved beyond the question of basic access. Attention is now focusing on the issue of what young people need to know about the technology that is, the forms of competence and understanding they need if they are going to use technology effectively and critically. [1]

As the convergence of media is growing, the boundaries between information and media are becoming increasingly blurred and hence we can no longer regard new media as a matter of just technology or information[1].

The major issue is the decreasing ability of a consumer to evaluate information and use it wisely, and this is mainly due to the rapid growth in the access of information thus the number of educated information consumers are decreasing.

The main problems in evaluating information lie in the difficulty of assessing the credibility and originality of information and the professional integrity of its presentation[2]. The boundary between traditional news that is information and user-generated content that is disinformation has also become increasingly blurred. Hence there is a need for the development of automated tools and technologies to assist the consumers in identifying the information out of disinformation by verifying, checking and filtering the content.

People who are information literate are good at "knowing when they have a need for information. identifying information needed to address a given problem or issue, finding needed information, evaluating the information, organizing the information, and using the information effectively to address the problem or issue at hand.[3]

2.c) Visualisation Techniques

The goal of visualization is to aid our understanding of data by leveraging the human visual system’s highly tuned ability to see patterns, spot trends, and identify outliers.
The challenge is to create effective and engaging visualizations that are appropriate to the data[4]. To create a visualisation number of factors are to be considered such as what questions are being answered by the visualisation, the type of data that is being used, what visual encoding should be used given a particular data which would again depend on various factors such as colour, size, shape and position. The challenge is that for any given data set the number of visual encodings and thus the space of possible visualization designs is extremely large[4].

The following are the types of visualisations that can be considered for this project:

Time-Series Data [5]

Sets of values that shift over time — or time-series data — are one of the most popular types of historical data. Time-changing trends are fundamental to other fields, such as finance (stock values, exchange rates), research (temperatures, emission levels, electrical potential) and public safety (crime rates). One also needs to equate a large number of time series concurrently and can pick from a variety of visualizations to do so.

Index Charts. [5]

For certain time-series data sources, raw values are less significant than conditional shifts. Consider buyers who are more involved in the pace of growth of the stock than their particular costs. Various stocks can have radically different reference values, which can be substantially similar to standardized stocks. The index map is an interactive line map that displays percentage adjustments to the time-series data set depending on the chosen index level. For eg, the picture in Figure 1a demonstrates the percentage increase in selected stock prices when bought in January 2005: one can see the rocky rise experienced by anyone who invested in Amazon, Apple or Google at that period.

Stacked Graphs. [5]

Many types of time-series data can be best interpreted as a whole. By piling region charts on top of each other, we arrive at a visible total of time-series values — a stacked graph. This type of graph (sometimes referred to as a stream graph) depicts aggregate patterns and often supports drill-down into a subset of individual series. Figure 1b shows the number of unemployed workers in the U.S. over the last decade, divided by industry. Although these charts have

proved to be common in recent years, they also have some major limitations. A stacked graph does not accept negative numbers and is meaningless to details that need not be summed up (temperatures, for example). Moreover, stacking may make it difficult to accurately interpret trends that lie atop other curves. Interactive search and filtering are often used to compensate for this problem.

Small Multiples. [5]

Instead of stacking, multiple time series can be plotted within the same axes as in the index chart. However, placing multiple series in the same space may produce overlapping curves that reduce readability. The alternate solution is to use tiny multiples: display each sequence in its own table. Figure 1c demonstrates again the number of unemployed employees, yet divided within each segment of the industry. We will now see both the general dynamics and the seasonal variations in each sector more specifically. When we're discussing time-series results, remember that tiny multiples can be designed for just about any sort of visualization: bar charts, pie charts, graphs, among others. This often produces a more effective visualization than trying to coerce all the data into a single plot.

Horizon Graphs. [5]

What happens when you try to equate multiple time series at once? Horizon graph is a methodology used to increase the data density of a time-series view while retaining detail. Find the five graphs listed in Figure 1d above. The first is a standard area chart with positive values for coloured blue and negative values for coloured red. The second graph "mirrors" negative values in the same region as positive values, twice the field graph data count. The third graph, the horizon graph, doubles the data density again by dividing the graph into bands and layering them to create a nested form. The effect is a map that retains the quality of the data but uses just a quarter of the room. Although the horizon

graph takes some time to learn, it has been found to be more effective than the standard plot when the chart sizes get quite small.

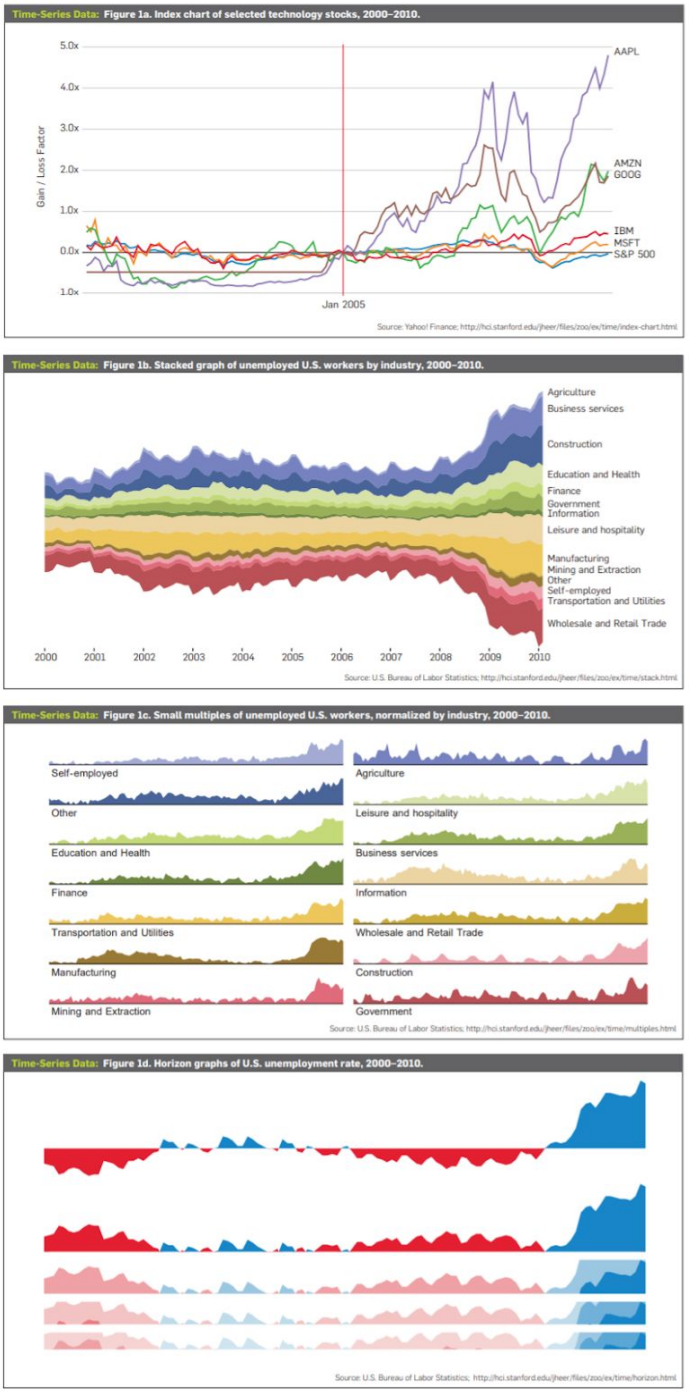

Fig4: Various types of graphs

Statistical Distributions [5]

Many visualizations were intended to demonstrate how a series of numbers is spread and thereby allow the observer to better understand the statistical properties of the results. Analysts also choose to adapt their data to mathematical models, either to check theories or forecast potential values, but an incorrect choice of model will contribute to inaccurate predictions. The exploratory data analysis is also an essential application of visualizations: to obtain insight into how data is processed to support data processing and modelling decisions. Popular strategies include a histogram that displays the distribution of values clustered into bins and a box-and-whisker plot that may express statistical features such as mean, median, quartile limit, or significant outliers. In addition, a variety of other methods exist to test the distribution and to analyze relations across different dimensions.

Stem-and-Leaf Plots. [5]

One alternative to the histogram is the stem-and-leaf plot to test a set of numbers. Usually, it connects numbers by the first significant digit, and then stacks the values inside each bin by the second significant digit. The minimalistic representation uses the data itself to depict the frequency spectrum, to remove the "information-empty" bars of the standard histogram bar map, and to allow one to determine both the overall spectrum and the contents of the bin. Figure 2a displays the distribution of success levels for staff performing crowdsourced activities on the Amazon Mechanical Turk. Note the various clusters: one cluster around a large degree of completion (99 % – 100 %); the other extreme is a cluster of Turks that accomplish just a few tasks (~10 %) in a group.

Q-Q Plots. [5]

While the histogram and the stem-and-leaf plot are typical instruments for determining the frequency distribution, the Q-Q (quantile-quantile) plot is a more efficient method. The Q-Q map contrasts two likelihood distributions by comparing their quantiles against each other. If the two are identical, the plotted

values would lie evenly along the central diagonal. If the two are linearly connected, the values should lie along the line again, but with different slopes and intercepts. Figure 2b displays the same details for Mechanical Turk involvement compared to three statistical distributions. Remember how the data contains three distinct components as opposed to uniform and regular (Gaussian) distributions: this implies that a mathematical model of three components might be more fitting, and in reality, we see in the final plot that the fitting mixture of three normal distributions is a stronger match. While powerful, the Q-Q plot has one apparent drawback in that its successful usage needs audiences to have some statistical knowledge.

SPLOM (Scatter Plot Matrix). [5]

Many visualization methods aim to reflect the relationship between several variables. Multivariate data exists often and becomes extremely challenging to depict, partially owing to the complexity of psychologically representing data across more than three dimensions. One strategy to solve this question is to use tiny multiple scattered plots showing a collection of pairwise relationships between variables, thereby generating a SPLOM (Scatter plot matrix). SPLOM enables visual analysis of the differences of each pair of variables. In Figure 2c, the scatter plot matrix is used to represent the characteristics of the car database, demonstrating the association between horsepower, weight, acceleration, and displacement. Additionally, interaction techniques such as brushing-and-linking—in which a selection of points on one graph highlights the same points on all the other graphs—can be used to explore patterns within the data.

Parallel Coordinates. [5]

As seen in Figure 2d, parallel coordinates) (take a specific approach to representation of multivariate results. Instead of graphing each pair of variables in two dimensions, we frequently map the data on parallel axes and then link

the related points to the row. That poly-line reflects a single row in the table, so cross-lines between measurements also suggest an opposite association. Reordering dimensions can assist with pattern-finding, as can dynamic database searching around one or more dimensions. Another feature of parallel coordinates is that they are fairly small, and several variables can be viewed concurrently.

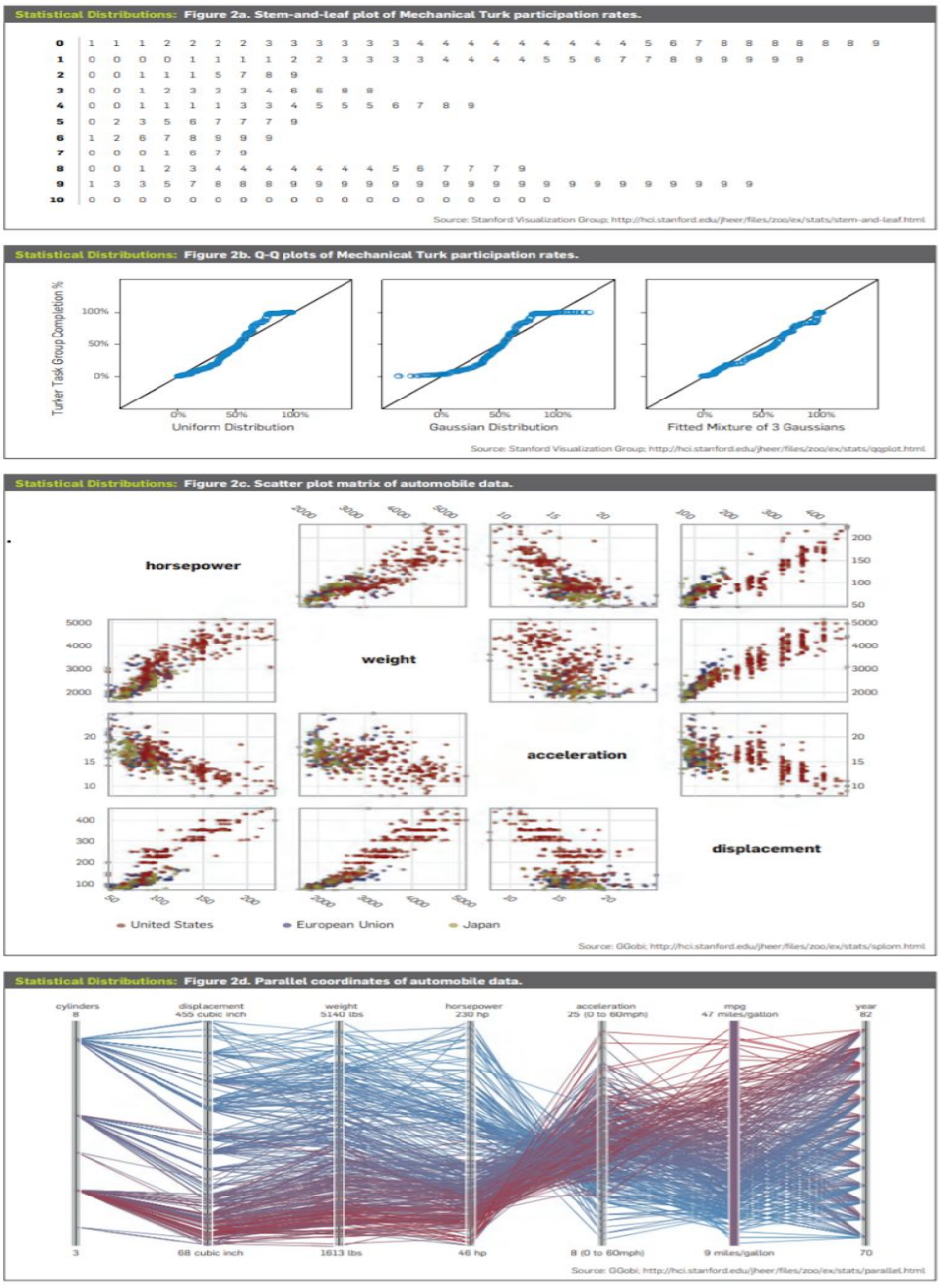

Fig5: various types of graphs 2

It is a very common technique for visualisation but in spite of its importance, little work has been done to study this visualization paradigm as a methodology in its own right. It is usually used in the ranking of search results and to show relationships among disparate entities [6]
Star Pattern is a type of Radial visualisation.

2.d) State of the Art Technologies

2.d.1) Interactive Storytelling to Teach News Literacy to Children[7]

This is used to improve news literacy amongst children usually between seven to nine years of age. Through various games and interactive storytelling using digital media, students of both school and colleges are injected with new literacy.

Knowing how to pick information and how to analyze it has become a crucial ability for everyone. In the case of teenagers, even though they do not ingest news of their own, they are sometimes subject to adult conversations on current affairs. If the disadvantaged community begins to grow, children will need further training for news reporting. Previous research has shown that news literacy (NL) education can lead to children's analytical thought, their sense of concern for global issues, and their civic involvement. NL can also contribute to reducing the perception of negative news material. [34]

By the age of five, most children have learned to distinguish fact from fiction and motive-action-consequence. Most of them are aware of the real existence of news reports, exhibiting more fear of upsetting news articles than terrifying fiction tales. Around the age of nine, children tend to be increasingly interested in how the media is created. Studies in the USA show that there are many NL lessons that can be offered to young children. Any of the concepts gained from the news that teachers have provided to elementary school-aged children include realistic skills in media, such as news composition, and discrepancies between news types and newspaper pages. In addition, several instructors have often come up with certain more abstract ideas, such as the distinction between

information and misinformation and freedom of speech problems. However, NL education is still not widespread in schools. When elementary school curricula contain any notion of reporting, they do so to a minimal degree. For this reason, most of the documented classroom interventions still depend heavily on the teacher's personal initiative and commitment. Currently, there are only a few digital games, mostly in English, that propose to teach NL to young children. Informal education of this sort may supplement the shortage of more comprehensive news education in classrooms. While there is some research that illustrates the importance and usefulness of these games at the college and secondary level, relatively little is understood regarding the usage of video games to educate elementary school children.

2.d.2) Storyfinder: Personalized Knowledge Base Construction and Management by Browsing the Web[8]

The web browser plugin listens and reacts to user events, initiates the analysis of a webpage, and provides a side pane view with the collected information. The server backend analyzes the webpage, extracts its metadata and stores it for later retrieval. The interactive webpage is embedded in the plugin’s side pane view and provides access to the newly gathered information. The Webpages which are visited by the user are analyzed by means of natural language processing components, and metadata are extracted in the form of named entities and keywords and stored for further reference.[8]

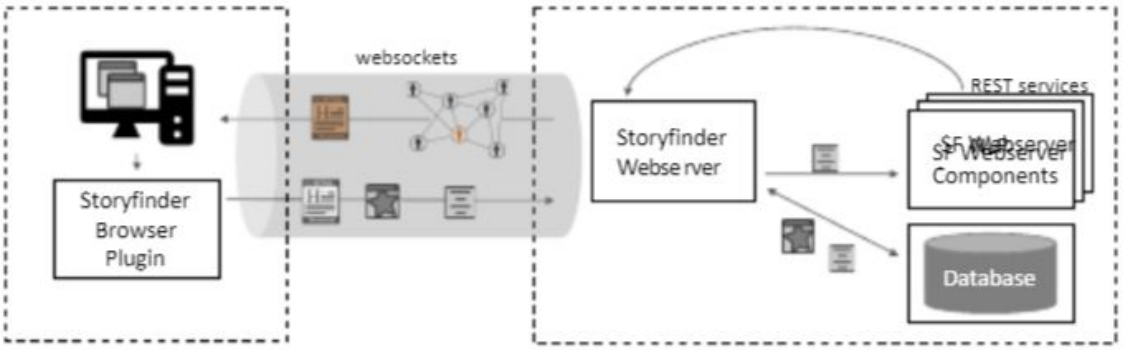

Fig6: Schema of Storyfinder's architecture and its components

Fig7: Screenshot of the default Storyfinder plugin view. A currently opened webpage is analyzed, the extracted entities are highlighted in an overlay, and rendered in a graph together with their relations in Storyfinder's interactive webpage, which is shown in a side panel of the browser.

Today, the internet is clearly the primary conduit of the information needs of society. Both news items or general truth, the internet, with its absolute

immeasurable speed of distributing new data and its enormous quantity of accessible knowledge, is the first option for information seekers. It is the user's right to read or miss a web page or a bookmark for later reference, but given that human memory may be unreliable, it is also the user's responsibility to maintain details tidy and readily available if a later reference is necessary. Instruments exist, such as concept maps, or mind maps among many others, which provide the necessary methodology and have been implemented in a multitude of prolific, computerized toolkits, which go beyond simple bookmarking.[35]

Their program comprises of three key components that will be described in more depth in the following sections:

1. The web browser plugin listens and reacts to user events, initiates the analysis of a webpage, and provides a side pane view with the collected information
2. The server backend analyzes the webpage, extracts its metadata and stores it for later retrieval
3. The interactive web page provides access to the newly gathered information and is embedded in the plugin’s side pane view.

2.d.3) Adaptive, Personalized Diversity for Visual Discovery[9]

When a person knows what they are looking for exactly that is if they have explicit intent then as a search query performs very well but when they are unsure of their intent then the performance of a search query is poor hence there is a need for a visual browsing system that emphasizes adaptive and personalized diversity in user experience. Such methods can be applied to any online system where the user seeks to discover content in the absence of an explicit search query.[9]

3) Theory and Design

a) Introduction

b) Definition of Problems and Solution

3.b.1) Addressing the task

3.b.2) Initial Ideas and possibilities

c) Resources

3.c.1) Hardware

3.c.2) Software

3.a) Introduction

This chapter discusses the concept of a question that defines the tasks at hand and then goes on to explore the initial ideas and suggestions to address such issues. This goes on to explore the concrete goals and priorities of the project. Finally, it adequately addresses the resources needed for this project.

3.b) Defining the problem

With the rapid growth in access to information, the ability of consumers to evaluate and use it wisely has become a key issue in creating educated information consumers. The main problems in evaluating information lie in the difficulty of assessing the credibility and originality of information and the professional integrity of its presentation. The boundaries between information and disinformation are becoming increasingly blurred.

3.b.1) Addressing the tasks

To build a machine learning model the following tasks are needed to be addressed:

1) Gathering Data
2) Preparing Data
3) Choosing a model
4) Training the model
5) Evaluation of the model
6) Tuning the parameters
7) Prediction

All the above steps are further discussed in details in the next chapter.

3.b.2) Ideas and Possibilities

The first task at hand was to come up with ideas to solve the problem of information literacy. After an extensive literature review, the idea is to design a machine learning model which will incorporate information literacy and build an environment that will provide consumers with a platform that will help segregate information and disinformation and then to implement the designed machine learning model through a pipeline of complex steps which included data gathering, data processing, data wrangling, selecting and training the classifier models.

3.c) Resources

3.c.1) Hardware

A computer workstation is the most important part of the development of this model, the specifications of the computer work station used are

Cpu- intel i5
Ram- 4 Gb
Graphic memory- Nvidia 820M
Os version - Ubuntu

3.c.2) Platform

Why Python?
Python offers concise and readable code. While complex algorithms and versatile workflows stand behind machine learning and AI, Python's simplicity allows developers to write reliable systems. Developers get to put all their effort into solving an ML problem instead of focusing on the technical nuances of the language.[38]

Additionally, Python is appealing to many developers as it's easy to learn. Python code is understandable by humans, which makes it easier to build models for machine learning.

Many programmers say that Python is more intuitive than other programming languages. Others point out the many frameworks, libraries, and extensions that simplify the implementation of different functionalities. It's generally accepted that Python is suitable for collaborative implementation when multiple developers are involved. Since Python is a general-purpose language, it can do a set of complex machine learning tasks and enable you to build prototypes quickly that allows you to test your product for machine learning purposes.[38]

4) Implementation

a) Introduction

b) Implementation

4.b.1) Data Identification

4.b.2) Data Gathering

4.b.3) Data Processing

4.b.4) Data Wrangling

4.b.5) Model Building

d) Challenges

4.a) Introduction

The subsections explain the implementation of the project. It will continue with a discussion on the tasks at hand, which is the specifications of the model, and then a further discussion will concentrate on what software was used for implementation and why the software was used and, in the end, the issues encountered during implementation will be addressed.

4.b) Implementation

The whole idea of the thesis is that false news is different from true news, and machine learning can detect this difference. This difference may be in any way that writing skills in a fake news article might be considerably lower than that of a true news article, or in other words, it might be less readable than true news, or the difference might be the writing style, etc.

4.b.1) Data Identification

But to build a machine learning model, first of all, we need an objectively classified data collection that can show us examples of false news and examples of actual news as given by professional fact-checkers. This implied, in effect, that the data would have URLs that were markers of the different news articles and some sort of scores that would show the quality of the material of those posts. Initially, a Washington Post Dataset was proposed, but some issues were found when dealing with that dataset, which is further addressed in subsection 4.c. After running through a variety of datasets, three datasets were identified that fulfilled the dataset criteria required to construct a machine learning model. The three databases are addressed in more depth in the next subsection.

4.b.2) Data Gathering

The first dataset is accessible from Github[10]. The dataset is a comma-separated CSV format with the columns below and it comprises all the true news URLs.

a) id - Unique identifier for each news
b) URL - Url of the article from the web that published that news
c) title - Title of the news article
d) tweet_ids - Tweet ids of tweets sharing the news. This field is a list of tweet ids separated by a tab.

The second dataset is accessible from Github[10]. The dataset is a comma-separated CSV format with the columns below and it comprises all the fake news URLs.

a) id - Unique identifier for each news
b) URL - Url of the article from the web that published that news
c) title - Title of the news article
d) tweet_ids - Tweet ids of tweets sharing the news. This field is a list of tweet ids separated by a tab.

For both of the aforementioned datasets, we don't need to use twwet_ids as well as title columns.

The third dataset is accessible on Ad Fontes Media[11]. The dataset is a comma-separated CSV format with the columns below and it comprises of both fake news as well as true news URLs.

a) Source- Source for each news
b) URL - Url of the article from the web that published that news
c) Bias - BIas score of the media house

d) Quality - Quality score of the pointed article

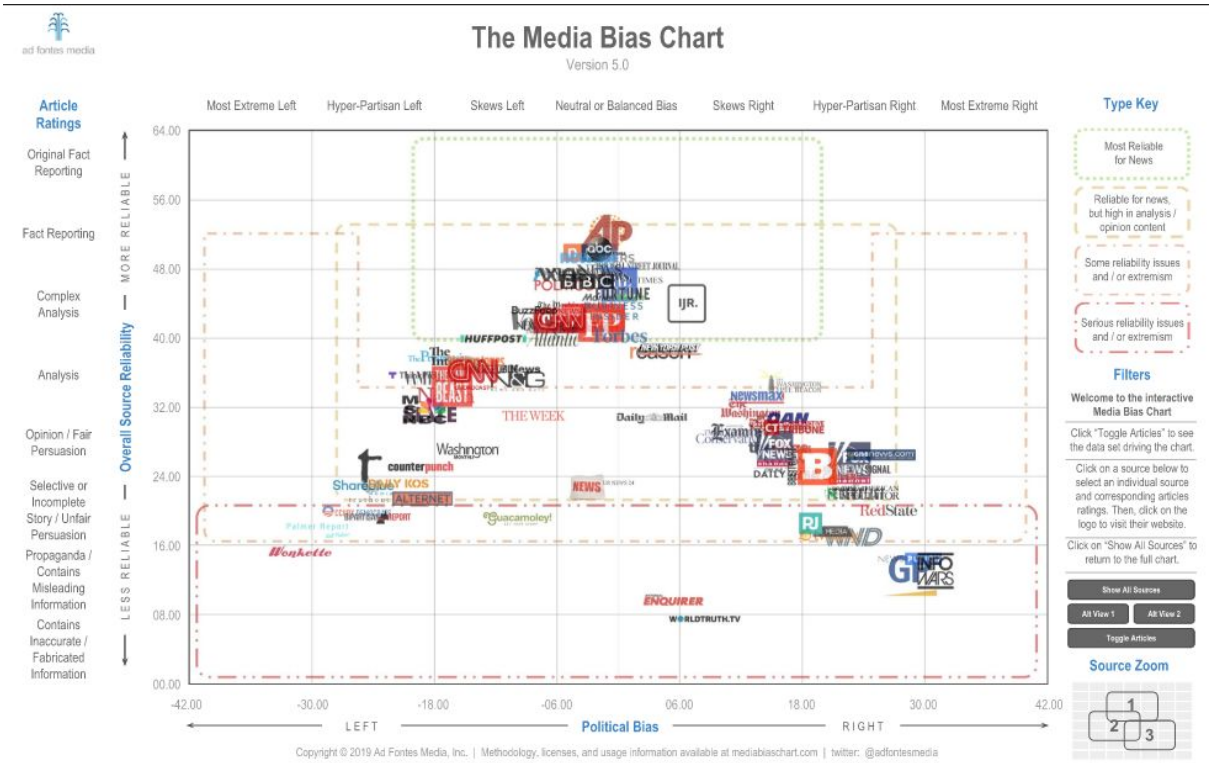

Fig: Media Bias Chart

4.b.3) Data Processing

Once the data has been obtained, the irregularities will be extracted from the databases and it is transformed into the required format.

In the Github datasets, several manual adjustments were made to the data sets to delete everything that was not a URL for example links to PDFs or word documents. The title and tweet_ids columns were omitted to form the dataset. After that, a new column was introduced in both databases, which effectively marked the data as false in the false data CSV file and true in the true data CSV file.

In the Ad Fontes Media dataset, the dataset was already in the correct shape, meaning that it can be used as it is.

After all the processing all three datasets were appended together and combined into one dataset.

4.b.4) Data Wrangling

Data wrangling, also referred to as data munging, is the method of transforming and mapping data from one "raw" data format to another format in order to make it more suitable and useful for a number of downstream uses, such as analytics[12].
The data wrangling process was the most time-consuming and complicated aspect of the project. It is further broken into three parts:

a) Reading our dataset consisting of scores and URLs.
b) Opening, reading and parsing the page pointed by the URLs.
c) Extracting and storing the parsed entities in text formats.

Collector Program (Collector.py), first imports and uses Pandas to read our URLs and scores in the CSV format. Then it calls the extractor program(Extractor.py) which in turn parses the articles pointed by the URLs and store them in text files, the working is discussed further in this section.

```
print("Ready to harvest Politifact data.")
input("[Enter to continue, Ctl+C to cancel]>>")
print("Reading URLs file")
df_csv = pd.read_csv("newsbot/politifact_data.csv", error_bad_lines=False, quotechar='"', thousands=',', low_memory=Fals
for index, row in df_csv.iterrows():
    print("Attempting URL: " + row['news_url'])
    if(ss.loadAddress(row['news_url'])):
        print("Loaded OK")

        if(len(ss.extractText)>500):
            ae = ArticleExample()
            ae.body_text = ss.extractText
            ae.origin_url = row['news_url']
            ae.origin_source = 'politifact data'
            ae.bias_score = 0
            ae.bias_class = 5
            ae.quality_score = row['score']
            ae.quality_class = row['class']
            ae.save()
            print("Saved, napping for 1...")
            time.sleep(1)
        else:
            print("**** This URL produced insufficient data.")
    else:
        print("**** Error on that URL ^^^^^")
```

Fig8: a snippet of collector Program

Pandas:

Pandas is a Python package that provides fast, flexible and expressive data structures designed to make working with "relative" or "labelled" data both easy and intuitive. It aims to be a key building block for Python's practical, real-world data analysis. In addition, it has a wider goal of being the most effective and scalable open-source data analysis/manipulation platform accessible in any language[13]. Pandas is well suited for many different kinds of data:[13]

- Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet
- Ordered and unordered (not necessarily fixed-frequency) time-series data.
- Arbitrary matrix data (homogeneously typed or heterogeneous) with row and column labels
- Any other form of observational/statistical data sets. The data actually need not be labelled at all to be placed into a pandas data structure.

But the actual work begins after the data files have been accessed, that is opening the URLs, reading the articles and then parsing the relevant entities. The biggest difficulty is clearly to isolate what is required and delete the

inconsistencies, that is, to only extract the actual English terms. The second challenge is to scale the extracted data in such a way that it is manageable on the PC which is only what is required to construct the model without losing efficiency and to do so we just need to save the root words, for example, if we consider the following words sing, sing, sing, sing, sing, we can stem all the four words down to just one-word sing.

Extractor program (extractor.py) is used to perform above-aforementioned tasks. Three libraries are used in the extractor program:

BeautifulSoup:

Beautiful Soup is a Python library for pulling data out of HTML and XML files.[14]Three features make it powerful:[15]

1. Beautiful Soup includes a few basic methods and Pythonic idioms for browsing, scanning, and changing a parse tree: a database dissecting toolkit and extracting what you need. It doesn't require a ton of coding to create an application
2. Beautiful Soup automatically converts incoming documents to Unicode and outgoing documents to UTF-8. You don't have to think about encodings unless the document doesn't specify an encoding and Beautiful Soup can't detect one. Then you just have to specify the original encoding.
3. Beautiful Soup sits on top of common Python parsers including lsml and htmlib5, enabling you to test out various parsing techniques or trade speed for flexibility.

urllib3:

urllib3 is a powerful, *sanity-friendly* HTTP client for Python which is used for loading web pages from the internet. urllib3 brings many critical features that are missing from the Python standard libraries:[16]

- Thread safety.

- Connection pooling.
- Client-side SSL/TLS verification.
- File uploads with multipart encoding.
- Helpers for retrying requests and dealing with HTTP redirects.
- Support for gzip, deflate, and brotli encoding.
- Proxy support for HTTP and SOCKS.
- 100% test coverage.

PorterStemmer:

The Porter stemming algorithm (or 'Porter stemmer') is a method to extract commoner morphological and inflexional ends from terms in English. Its main use is as part of a standardization process that is usually carried out when setting up information retrieval systems.[17]

```
from bs4 import BeautifulSoup
from bs4.element import Comment
from urllib3.exceptions import HTTPError
import urllib3
import re, string
import json
import html
from io import StringIO
from nltk.stem import PorterStemmer
```

Fig9: a code snippet showing the three libraries

Now, after importing the three libraries, configure an English dictionary that can be found on Github, which will help in comparing terms such that only real English terms are extracted and a list comprising the extracted terms is initialized. After the extractor program is all set up, the collector program is executed which will call the extractor program and end up with a full set of real English stemmed words for every URL. So now the dataset has stemmed real English words associated with each URL and a quality score which was already

there in our initial dataset. With that we're finished with the data wrangling phase.

4.b.5) Model Building

Next step in the process is building a model. This process is divided into four steps.

a) Constructing a Matrix
b) Building a list of extracted words associated with a primary key
c) Classifier Models
d) Validating the Models

The design of a matrix is such that each row in the matrix will represent one URL/article, each row will be made up of 0’s and 1’s wherein 0 indicates that the word is not present in the article and 1 indicates that a word is present in the article under consideration. For which each word needs to be uniquely identified and hence a primary key is assigned to every extracted word in the database. Yet we need a fresh list of just extracted terms to which a primary key can be assigned.

A new list containing extracted words is constructed using the list build program(list_build.py) so as to make the program more efficient as for every next URL/article the words in that article are first compared with the words in the dynamic list and since dynamic list has only extracted words, the comparison will take less time relatively as compared to comparison with the dictionary every time. Since the list is dynamic, any extracted term that has not been a part list before will be added to the list. This list is called recursively for every URL. Now that a list of extracted words is created, every word in that list is assigned with a primary key which is used to identify them uniquely.

Next step is to classify the articles. For the classification the difference between the upper bound and lower bound of the quality scores was divided by 4, for instance if the lower bound is 0 and upper bound is 100 then 0-25 will be class 1 which is False news, 25-50 will be class 2 which is mostly fake, 50-75 will be class 3 which is mostly true and 75 above will be class 4 which is true. To classify the articles into one of these four classes is the real task for the classifier model.

The obtained matrix is of size 21000 x 2450 wherein 2450 is the total number URLs/article considered for training data and 21000 words were extracted from the articles, along with the quality scores of those 2450 articles are used to train and test the model.

To build a classifier model sklearn library is imported.

Sklearn:

Scikit-learn is an open-source machine learning library that supports supervised and unsupervised learning. It also provides various tools for model fitting, data preprocessing, model selection and evaluation, and many other utilities[22]. Its critical features include

- Simple and efficient tools for predictive data analysis
- Built on NumPy, SciPy, and matplotlib

Using sklearn several classifier models can be imported and the best part is it is very easy and convenient to use any mode. For this project Support Vector Machine Classifier and Logistic Regression were implemented.

Support Vector Machine:

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection[23].

The advantages of support vector machines are:[23]

- Effective in high dimensional spaces.
- Still effective in cases where the number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector machines include:[23]

- If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial.
- SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation (see Scores and probabilities, below).

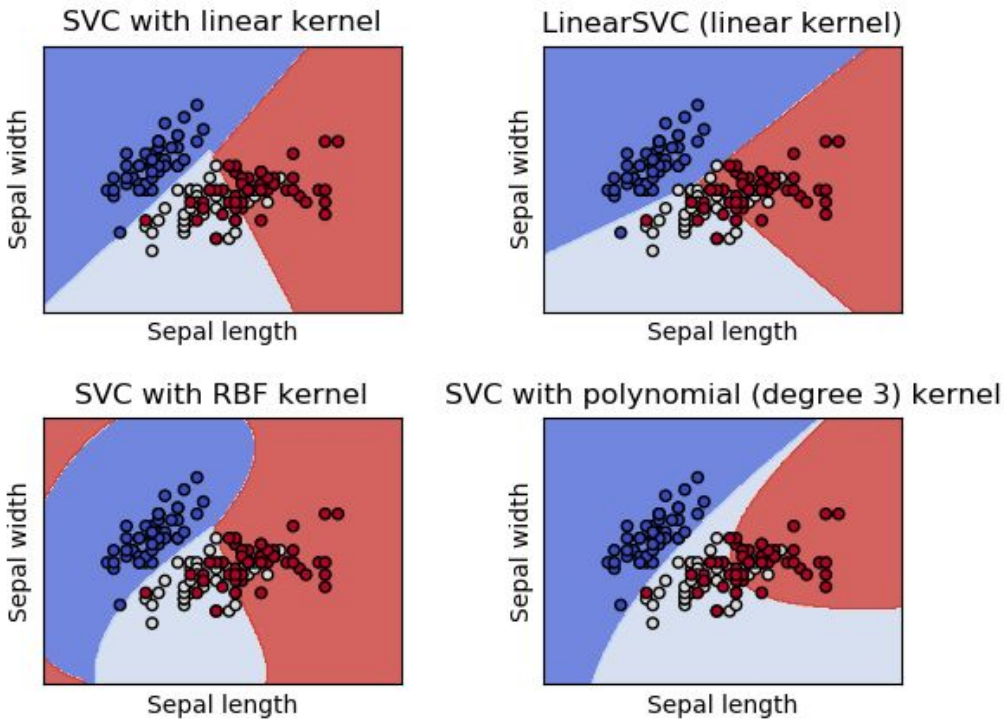

Fig10: SVM Classifier

A support vector machine constructs a hyper-plane or set of hyper-planes in a high or infinite-dimensional space, which can be used for classification, regression or other tasks. Intuitively, a good separation is achieved by the hyper-plane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier[25].

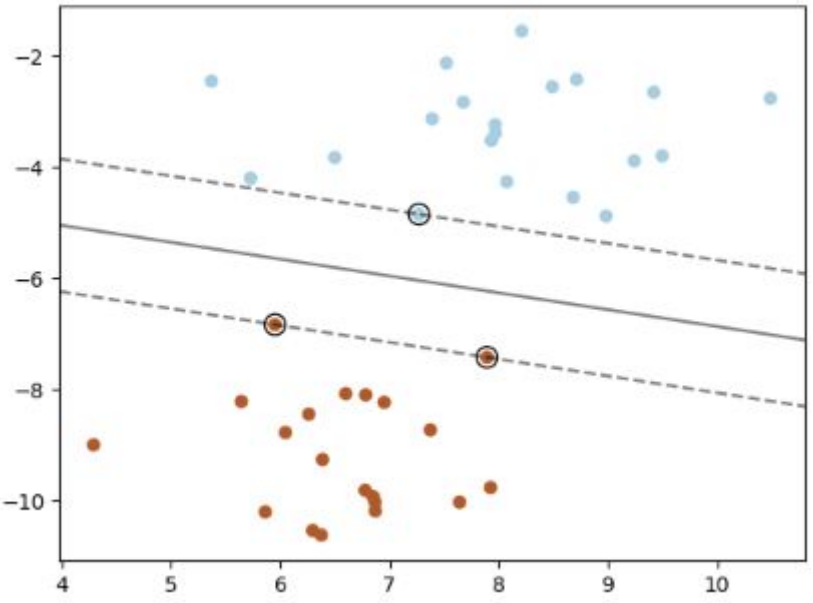

Fig11: SVM hyperplanes

The prediction for svm is:

$$h_\theta(x) = sign(\theta^T x) \tag{1}$$

for parameter vector θ and input x. Also, the cost function is:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \max(0, 1 - y^{(i)}\theta^T x^{(i)}) + \lambda\theta^T\theta \tag{2}$$

and the gradient of the cost is a vector where the jth element is:

$$2\lambda\theta_j - \frac{1}{m}\sum_{i=1}^{m} y^{(i)}x_i^{(i)}\mathbb{1}(y^{(i)}\theta^T x^{(i)} \leq 1) \tag{3}$$

Logistic Regression:

Logistic regression, despite its name, is a linear model for classification rather than regression. Logistic regression is also known in the literature as logit regression, maximum-entropy classification (MaxEnt) or the log-linear classifier. In this model, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function[24].

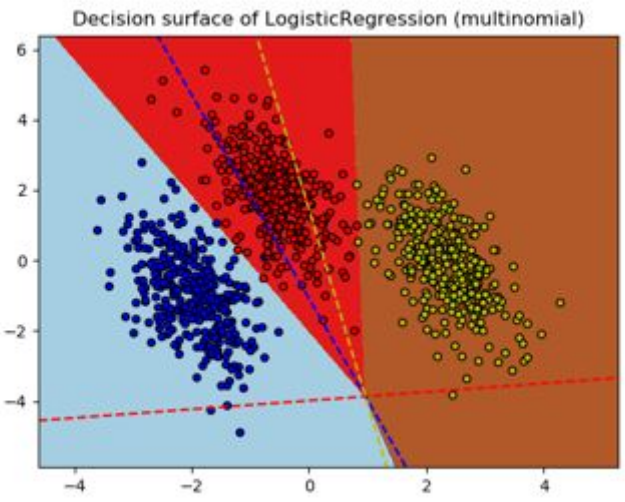

Fig12: Decision Surface of Logistic Regression

The prediction for Logistic Regression is :

$$h_\theta(x) = sign(\theta^T x) \tag{4}$$

for parameter vector θ and input x. Also, the cost function is:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \log(1 + e^{-y^{(i)}\theta^T x^{(i)}}) \tag{5}$$

and the gradient of the cost is a vector where the jth element is:

$$\frac{\partial}{\partial\theta_j} J(\theta) = \frac{1}{m}\sum_{i=1}^{m} -y^{(i)} x_j^{(i)} \frac{e^{-y^{(i)}\theta^T x^{(i)}}}{1+e^{-y^{(i)}\theta^T x^{(i)}}} \tag{6}$$

The final stage in the process is to validate the model by evaluating several performance parameters.

Confusion Matrix:

It is a 4 x 4 matrix, with rows reflecting the predicted class and column reflecting the real class. Therefore, the intersection of row and column indicates where the real class was the same as the predicted class, for instance, if the value of 4th row and 4th column is 50 then there were 50 instances where the article was true and it was classified true. Whereas if the value of 1st row and 2nd column is 5 this implies that there were 5 instances where predicted class was 1that is false but they were actually in class 2.

For both SVM and Logistic Regression, it is observed that the majority of the classification were true or they were classified one class above or below their actual class.

Accuracy:

It is the ratio of a number of correct predictions to the total number of input samples. The accuracy for support vector machine classifiers was 64% whereas the accuracy for logistic regression was 57%. Considering the equally likely probability of each class which is 25%, it can be clearly observed that both the models are more than twice as accurate.

Hence it can be validated that the performance of both the models is satisfactory. After validation both models are executed, trained and saved. A simple command-line interface is used to test the model which lets the user input the article link and a probability estimate of all four classes on those articles is printed. Evaluation of the model is discussed in details in chapter 5.

```
print("Statistical tests...")
print("***************")
print("Accuracy score: " + str(accuracy_score(predictions, y_test)))
print("Confusion Matrix: ")
print(confusion_matrix(predictions, y_test))
print("Classification report: ")
print(classification_report(predictions, y_test))
print("***************")
print("Regression based: ")
rSq = r2_score(y_test, predictions)
expVariance = explained_variance_score(y_test, predictions)
maxErr = max_error(y_test, predictions)
mae = mean_absolute_error(y_test, predictions)
print("R^2: " + str(rSq))
print("Explained variance: " + str(expVariance))
print("Max Error: " + str(maxErr))
print("Mean absolute Error: " + str(mae))

exit(0)
```

Fig13: a code snippet showing validation parameters

4.c) Challenges

a) Problem Formulation:

It's a crucial thing to worry about when you start developing a machine learning model. It is very important to consider what you really intend to achieve and therefore take a very straightforward path to develop the model.

b) Data Engineering:

It is the most difficult challenge faced while developing any machine learning model. When in the process of identifying and gathering data, a Washington post dataset was initially used, but in the end, it was actually not feasible to use that dataset to construct a model. Several issues were faced with that dataset, the first of which was the sheer scale of the data that was around 7 Gb in the form of a JSON format. It's not feasible to deal with that massive file on the local computer, because it will need a really

large RAM. To deal with the dataset, it was split into many files of smaller sizes, which made it easier to deal with the details but gave rise to a new issue. The division of data made all the smaller data very inconsistent with each other because there was no common start tag and end tag for each file. That is a file that can start with any tag and finish with another that will make it nearly difficult to parse. To solve these issues, google colab was used, which is a cloud application provided by Google. After accessing the file using colab, two significant shortcomings were found in the dataset first of which there was no quality score of some type and thus it was impossible to classify the articles in the four classes discussed above, which are Class 1 – false news, Class 2 – mostly false, Class 3 – mostly real or Class 4 – true. The second shortcoming was that it had only around 390 posts, which are too limited to train an effective machine learning model.

Wrangling of data was another major challenge encountered during building this model. There were several problems encountered which were, reading the dataset, opening the URLs, reading those URLs, extracting the readable content of the articles and then storing the extracted words in the dataset with the associated URLs and the quality scores.

c) Scalability:

Scalability is also a topic of concern. As mentioned in subsection (a) Washington post, the dataset was too big to cope with ordinary PCs. It is also important to scale down the data set in such a way that you satisfy the specifications of the model and at the same time do not sacrifice performance, such that it can also be handled on a normal PC. In this experiment, to scale down the sample, only real English terms were extracted from the documents by matching them with terms in the existing English dictionary. Secondly, words are stemmed that is only root words were extracted for instance run, ran would be stemmed down to

just run, due to this instead of ending up with 50000 words we ended up with only 20000 words which are less than half.

d) Time Complexity:

The current model takes up to 4 hours to run, which, if we talk about commercial aspects, can not be used commercially even if the interface is ready because it is too slow. Sure, hardware plays a role, but there's no question that the algorithm should be more time-efficient.

e) Choosing the appropriate Visualisation:

The goal is to construct efficient and interactive visualizations that are suitable for the data[4]. In order to construct a visualization, a variety of variables must be weighed, such as what questions are asked by the visualization, the form of data that is being utilized, what graphic representation would be used, provided the specific data that will again rely on multiple considerations such as colour, height, shape and location.

But more than a challenge it turned out to be more of a shortcoming for this project, which is discussed further in the limitations section in the next chapter.

f) User Interface:

Developing a user interface was again a big challenge as well as one of the shortcomings. If this model is to be deployed commercially a user interface is a must. It might be in the form of a browser plugin or some application, but the bottom line being there should be something that a user can interact with.

5) Evaluation

a) Introduction

b) Evaluation of performance

c) Key Findings

d) Limitations

5.a) Introduction

This segment would address the assessment and review of this framework. It will start by discussing the performance of the application on different issues and then moves on to discuss input and recommendations from the consumers as well as my final thoughts on it. In the end, the key findings of this project are addressed.

5.b) Evaluation of Performance

To assess the performance of this model, a data set composed of URLs that point to news articles and a label which basically labels all the URLs as true, false or mixture, would be tested upon. The classification result of this model will then be compared to the actual label of that URL to verify if the prediction was correct or incorrect.

49	mixture	https://w	Home Dep	Politics	06-Jun-16	Home Dep	http://ww	3	TRUE	No Error
50	mixture	https://w	George W	Politics	11-Oct-16	The G.W. I	https://w	3	TRUE	No Error
51	mixture	https://w	Donald Tr	Politicians	########	Donald Tr	http://ww	3	TRUE	No Error
52	FALSE	https://w	Clint East	Fake New	########	Clint East	https://ar	1	FALSE	No Error
53	FALSE	https://w	Student F	Fauxtogra	########	A student	http://ww	1	TRUE	No Error
54	FALSE	https://w	Lead in Fli	Medical	########	Videos of	http://ww	1	TRUE	No Error
55	mostly fal	https://w	Are Non-C	Ballot Box	########	Green Car	http://ww	3	TRUE	No Error
56	TRUE	https://w	Global Sea	Science	########	Daily sea i	http://ngc	3	FALSE	No Error
57	FALSE	https://w	Did a Stud	Ballot Box	########	An acader	http://ww	3	FALSE	No Error
58	FALSE	https://w	Shelby To	Crime	########	A 25-year-	http://pec	1	TRUE	No Error

Fig14: a snippet of the dataset used for testing

To compare the results I have chosen ten random URLs from the test dataset as it would be impossible to show results for every URL although that would be more accurate.

1) https://www.snopes.com/fact-check/buzz-aldrin-tweeted-we-are-all-in-danger-it-is-evil-itself/ - **False**

Fig15: Results for SVM Model

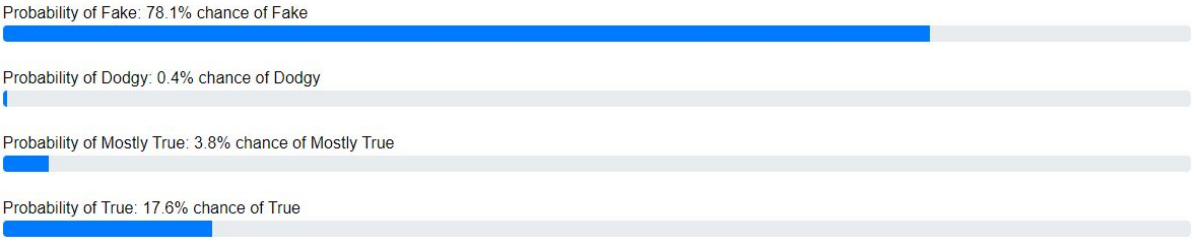

Fig16: results for Logistic Regression Model

This link is labelled false in the dataset. The SVM model predicted that there is a 35.4% chance that this article is fake which is correct but it is not as precise whereas logistic regression performs much better and it correctly predicts that this news article is fake by 78.1%.

Svm- 1/1

LR-1/1

2) https://www.snopes.com/fact-check/trump-taps-outspoken-climate-denier-to-oversee-epa-transition-team/ - **True**

Fig17: SVM Prediction

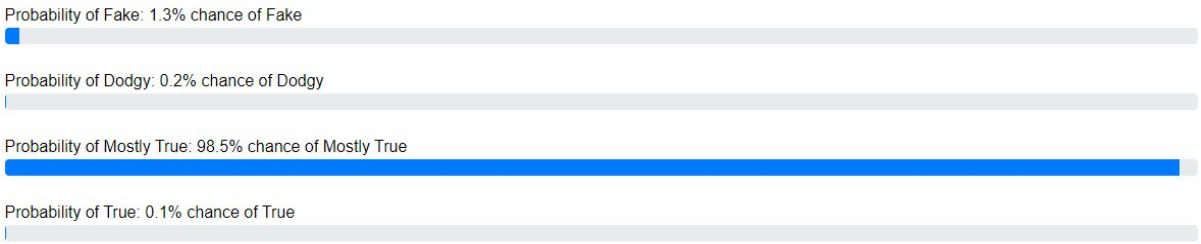

Fig18: LR prediction

Svm predicts 18.4 % that the article is true however it also predicts that there is 58.3% of the chance that article is mostly true and hence if we consider both the numbers we can say it is highly likely that this article is actually true. LR predicts that 98.% chance that this article is mostly true which can again be considered a correct prediction but not a precise one.

SVM- 2/2

LR-2/2

3) https://www.snopes.com/fact-check/weather-report-sweden-iraq/ - False(Mixture)

Fig19: SVM

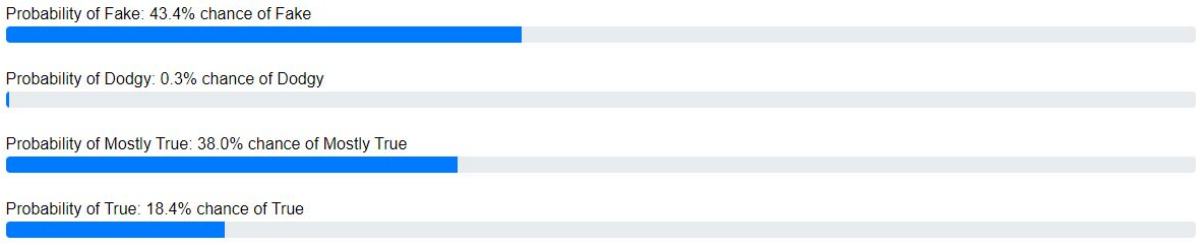

Fig20: Lr

Svm predicts that there is 36.3 % of the chance that this article is false which is correct as this article is labelled mostly false. Lr also predicts that 43.4% chances

that this article is false which is again correct. However, we can say that both the cases they weren't very precise.

SVM - 3/3

LR- 3/3

4) https://www.snopes.com/fact-check/peanut-butter-cups/ - False (Mixture)

Fig21: svm

Fig22: LR

Svm predicts that there are 56.7% chances that this article is mostly true which is an incorrect prediction. LR predicts that this article is 99% mostly true which is incorrect again.

SVM- 3/4

LR- 3/4

5) https://www.snopes.com/fact-check/australia-teen-sea-lice-fleas/ - True (Mixture)

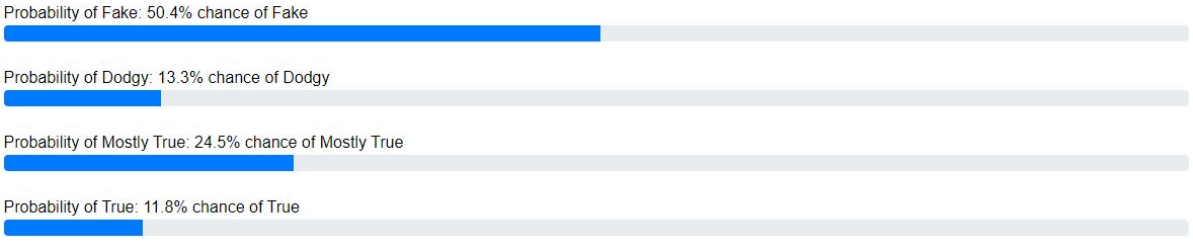

Fig23 : svm

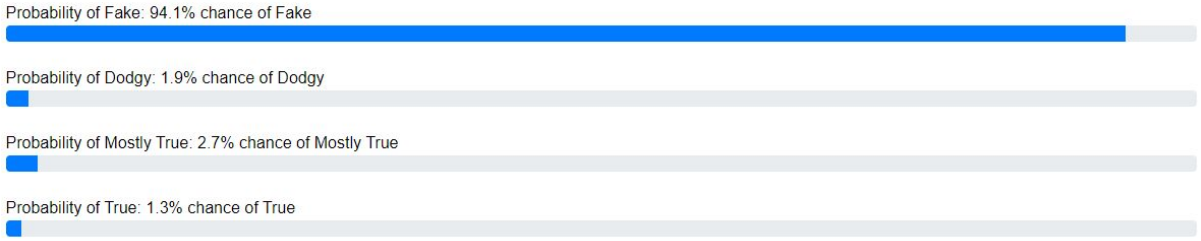

Fig24: LR

Svm predicts that 50.4% chances that this article is fake which is an incorrect prediction. LR predicts that there are 94% chances that this article is fake which is incorrect as well.

SVM-3/5

LR-3/5

6) https://www.snopes.com/fact-check/aunt-jemima-pancakes-recall/ - True (Mixture)

Fig25: svm

Fig26: LR

Svm predicts that there is a 45% chance that this article is fake which is incorrect. LR predicts that there are 59% chances that it is fake which is incorrect as well.
SVM- 3/6
LR-3/6

7) https://www.snopes.com/fact-check/burglars-return-stolen-computers/ - True

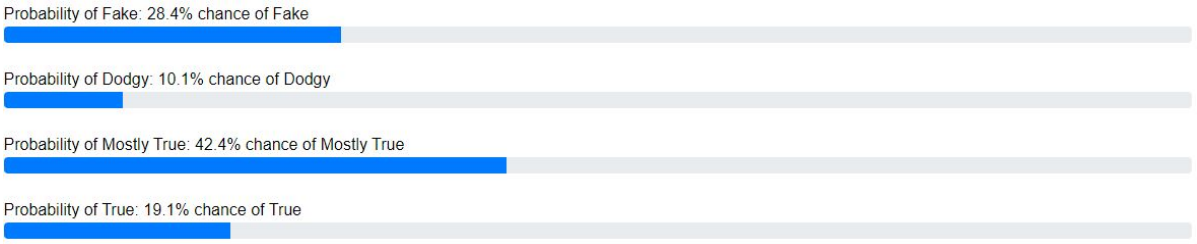

Fig27: svm

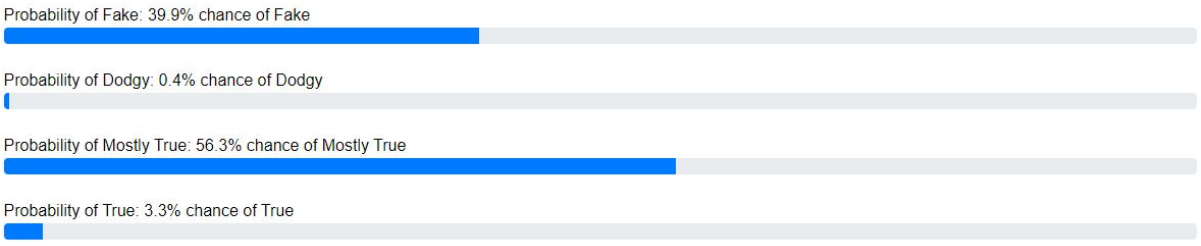

Fig28: LR

SVM predicts that there is a 42% chance that this article is mostly true and 20% that it is actually true, considering them together we can say that this was a correct prediction but not precise. LR predicts that there is a 56 % chance that

this article is mostly true and 3% chance that it is actually true, hence considering the two together we can say this is a correct prediction.
SVM- 4/7
LR-4/7

8) https://www.snopes.com/fact-check/jeff-rothschild-third-world-war/ - False

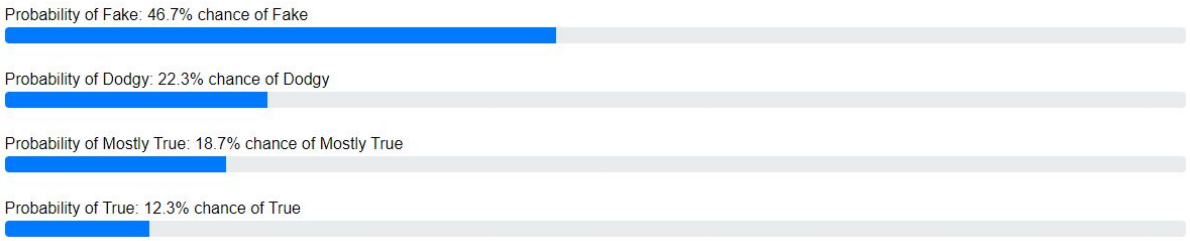

Fig29: svm

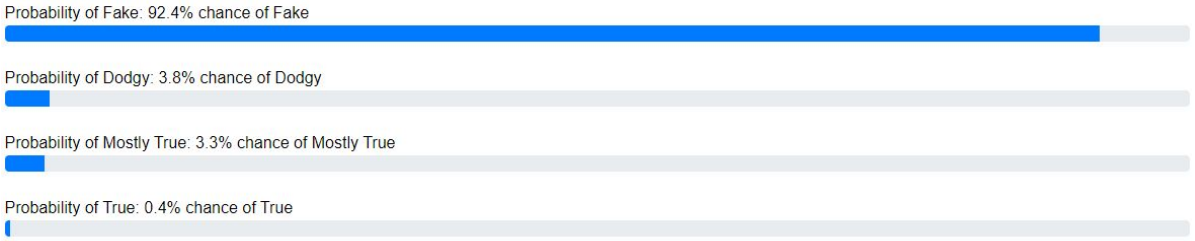

Fig30: LR

Svm predicts that there is a 47 % chance that this article is false which is a correct prediction. LR predicts that there is 92.4% chance that this is fake which is a correct prediction.
SVM-5/8
LR-5/8

9) https://www.snopes.com/fact-check/un-backs-secret-obama-takeover-of-police/ - False

Fig31: svm

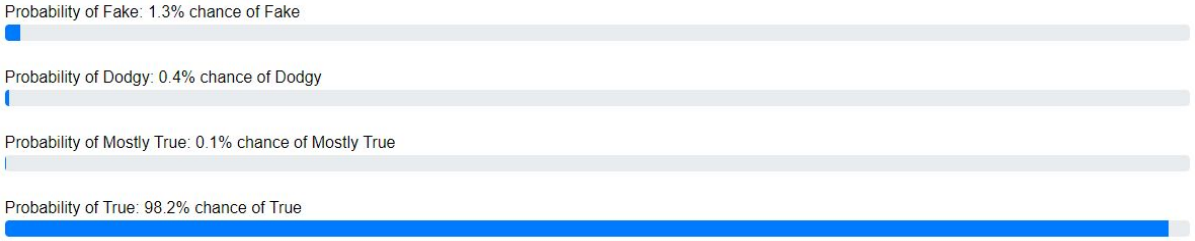

Fig32: LR

Svm predicts that this article is most likely to be true which is an incorrect prediction. LR predicts that this article is actually true which is again incorrect prediction.

SVM- 5/9

LR-5/9

10) https://www.snopes.com/fact-check/bill-maher-recession-get-rid-trump/ - True

Fig33: svm

Fig34: LR

Svm and LR both predict that this article is fake which is an incorrect prediction.
So the final scores are:
SVM-5/10
LR-5/10

5.c) Key Findings

From the above evaluations, we will note that the accuracy of both SVM and Logistic Regression is 50 per cent, and if we compare it with the accuracy that we have achieved when training a model that is 64% for SVM and 57% for Logistic Regression, this number is not poor. Model is expected to be less accurate than what we have measured because while testing the dataset is split into 80% and 20%, where 80% is the train data and 20% is the test data, all of which are part of the same dataset and thus the estimated accuracy is bound to be higher.

The other explanation for the loss of accuracy is that such URLs under consideration are all from the same source as Snopes, which was also not part of the training data set. Also, we've only evaluated 10 examples here but, in order to get a better understanding of the efficiency of the pattern, it requires to be tested in more instances.

Another interesting observation is that either the expected class was right or that the algorithms classified articles in 1 class above or below, for example, an article that is meant to be in Class 4 and is true is classified in Class 3 that is

mostly true. The classification gap between classes was 2 in just 2 out of 10 instances. It indicates that the performance of the model has been optimal.

In general, after testing the model on more instances it was observed that the percentage of correct prediction when it tested on a true news article is more than the percentage of correct prediction when it is tested on a false news article. It may be assumed that the model is slightly better in predicting the true news than the false news.

An interesting finding is that both SVM and LR have struggled to forecast correctly in the same cases, which clearly indicates that the accuracy of both models is identical and that one model may be used to double-check the prediction of the other models.

5.d) Limitations

a) **Visualisation:**

This project started with an idea of information literacy and why is essential. It was discussed that visual narratives can be used to integrate information literacy in the general working environment of a consumer. Star graph was the chosen visualisation which could have fulfilled the purpose. The concept was that axes of a star graph might represent sever aspects such as readability, originality, quality of content if it's real or fake, etc. However, only one of those axes could be implemented by the end. Hence this is definitely one of the shortcomings of this project.

b) **User Interface:**

A user interface is a must for an application of such kind because this is something which is to be used by a consumer at the front end as well. Hence a model of such type should not only have a back end but also a

front end with which a consumer can interact with. This interface can be a web browser plugin, or an extension or simply an application.

Apart from the two there are some general shortcomings of the project, the accuracy of the model is approximately 50% - 60% which is not bad but if we talk realistically it should be somewhere around 80%-85% when the model is deployed commercially. Hence there is a lot of room for improvement in the model in terms of performance.

The training data used only has around 2450 articles, which for real-world application is too fewer data to train on, to build a good model, it should be trained on a lot more instances but due to hardware limitations that wasn't possible at the moment.

6) Conclusion and future work

The basis of the research was a simple idea that people who are information literate are good at knowing when they have a need for information, identifying information needed to address a given problem or issue, finding needed information, evaluating the information, organizing the information, and using the information effectively to address the problem or issue at hand.

To Address this task a Machine learning model was designed which will incorporate information literacy and build an environment that will provide consumers with a platform that will help segregate information and disinformation.

The model was implemented through a pipeline of complex steps which included data gathering, data processing, data wrangling, selecting and training the classifier models and evaluation of those models. There Were several problems encountered during implementation especially during the data engineering process. Two classification models were implemented SVM and LR, which basically classified an article in one of the four classes, that is true, fake, mostly true and mostly fake.

The model was evaluated on a test dataset obtained from Snopes. The performance of the model was satisfactory as the accuracy of correct prediction was 50% for both SVM and LR though it is observed that the model is slightly better in the prediction of true news. There were few limitations with the models, it should have more visual elements as well as a user interface.

6.a) Future Work

There are certainly various aspects of the project that can be worked upon in future. First and the foremost thing will be to test it with consumers and get the user feedback which ideally would be done but due to the ongoing unavoidable

circumstances, it was not possible to test it with users. User feedback helps the most in developing the model further as the requirements change accordingly, for instance, a developer might like the bright interface of the application but it might not work well with the user, hence it's very important to get user feedback.

More visual elements should be added. The model should be developed further so that it does not evaluate one aspect, that is if a piece of news is true or fake but also other aspects of the news such as its originality, its readability, bias score, the quality of content and so on. Then those aspects should be represented visually using the best suited visual technologies such as star graph.

Any model of such type needs an interface with which a user can interact with so one of the future objectives is also to create a user interface which can be easily accessed and incorporated in the web environment so as to be used by the consumers.

7) References and List of figure

7.a) References

[1] Buckingham D. (2010) Defining Digital Literacy. In: Bachmair B. (eds) Medienbildung in neuen Kulturräumen. VS Verlag für Sozialwissenschaften

[2] Eshet-Alkalai, Yoram. (2004). Digital Literacy: A Conceptual Framework for Survival Skills in the Digital Era. Journal of Educational Multimedia and Hypermedia. 13.

[3] http://magpifellows.pbworks.com/f/21st+Century+Learning.pdf (visited on 1st November 2019)

[4] Jeffrey Heer, Michael Bostock, and Vadim Ogievetsky. 2010. A tour through the visualization zoo. Commun. ACM 53, 6 (June 2010), 59–67.

[5] I. Herman, G. Melancon and M. S. Marshall, "Graph visualization and navigation in information visualization: A survey," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 6, no. 1, pp. 24-43, Jan.-March 2000.

[6] G. M. Draper, Y. Livnat and R. F. Riesenfeld, "A Survey of Radial Methods for Information Visualization," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 5, pp. 759-776, Sept.-Oct. 2009.

[7] Campos, Ioli. (2017). Interactive Storytelling to Teach News Literacy to Children. 347-350. 10.1007/978-3-319-71027-3_40.

[8] Remus, Steffen & Kaufmann, Manuel & Ballweg, K. & Landesberger, Tatiana & Biemann, Chris. (2017). Storyfinder: Personalized Knowledge Base Construction and Management by Browsing the Web.

[9] Choon Hui Teo, Houssam Nassif, Daniel Hill, Sriram Srinivasan, Mitchell Goodman, Vijai Mohan, and S.V.N. Vishwanathan. 2016. Adaptive, Personalized Diversity for Visual Discovery. In Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16).

[10] https://github.com/KaiDMML/FakeNewsNet/tree/master/dataset

[11] https://www.adfontesmedia.com/?v=402f03a963ba

[12] https://en.wikipedia.org/wiki/Data_wrangling

[13] https://pandas.pydata.org/docs/getting_started/overview.html

[14] https://www.crummy.com/software/BeautifulSoup/bs4/doc/#

[15] https://www.crummy.com/software/BeautifulSoup/

[16] https://pypi.org/project/urllib3/

[17] https://tartarus.org/martin/PorterStemmer/

[18] http://www.therightinformation.org/realrelevant-whatis/

[19] http://www.therightinformation.org/realrelevant-importanceof/

[20] Virkus, Sirje. (2003). Information literacy in Europe: A literature review. Information Research: an international electronic journal. 8.

[21] Bruce, C.S. (2002). Information literacy as a catalyst for educational change: a background paper. White Paper prepared for UNESCO, the U.S. National Commission on Libraries and Information Science, and the National Forum on

Information Literacy, for use at the Information Literacy Meeting of Experts, Prague, The Czech Republic.

[22] https://scikit-learn.org/stable/index.html

[23] https://scikit-learn.org/stable/modules/svm.html

[24]https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

[25]https://scikit-learn.org/stable/modules/svm.html#svm-mathematical-formulation

[26] Association of College and Research Libraries. (2000). Information Literacy Competency Standards for Higher Education. Chicago: ACRL. Retrieved 11 July 2003 from http://www.ala.org/acrl/ilcomstan.html

[27] American Library Association. Presidential Committee on Information Literacy (1989). Final Report. Chicago: American Library Association.

[28] Alava, S. (1999). Mediation(s) et metier d'etudiant. Bulletin des Bibliotheques de France, 44(1), 8-15.

[29] Alvestrand, V. (2003). Support for info literacy certificate grows. Information World, 191, 1.

[30] Anttiroiko, A.-V., Lintilä, L. & Savolainen, R. (2001). Information society competencies of managers: conceptual considerations, In: E. Pantzar, R. Savolainen & P. Tynjälä, eds. In search for a human-centred information society. (pp. 27-57). Tampere: Tampere University Press.

[31] Armstrong, C.J., Lonsdale, R. E., Stoker, D.A. & Urquhart, C.J. (2000). JUSTEIS: JISC usage surveys: trends in electronic information services. Aberystwyth: University of Wales. Retrieved 10 January 2003 from http://www.dil.aber.ac.uk/dils/Research/Justeis/cyc1rep0.htm

[32] Audunson, R. & Nordlie, R. (2003). Information literacy: the case or non-

case of Norway? Library Review, 52(7). (Fothcoming)

[33] Audunson, R., Hansson, J., Nagy, A., Ulvik, S. & Westheim, I. (2001). Information literacy, public libraries and local community engagement are a European project on the potential of public libraries as promoters of local participation and viable local communities in metropolitan areas. In: Nordic seminar on public library research, 10-11 December, Copenhagen. (pp. 15-23) Copenhagen: Royal School of Library and Information Science.

[34] Bainton, T. (2001). Information literacy and academic libraries: the SCONUL approach. Proceedings of the 67th IFLA Council and General Conference, August 16-25, 2001. The Hague: International Federation of Library Associations. Retrieved 10 January 2003 from http://www.ifla.org/IV/ifla67/papers/016-126e.pdf

[35] Basili, C. ENIL network. Retrieved 10 January 2003 http://www.isrds.rm.cnr.it/personale/~basili/EnIL/index.html

[36] Bawden, D. (2001). Information and digital literacies: a review of concepts. Journal of Documentation, 57(2), 218-259.

[37] Bawden, D. & Robinson, L. (2001). Training for information literacy: diverse approaches. Proceedings of the International Online Information Meeting, London, 4-6 December 2001. (pp.87-90). Oxford: Learned Information Europe Ltd

[38] https://towardsdatascience.com/8-reasons-why-python-is-good-for-artificial-intelligence-and-machine-learning-4a23f6bed2e6

7.b) List of figures

1) Fig1:evolution of media
2) Fig2: info lit pyramid