# High-Level Design Report

VERITAS

**Team Members:**

- Hanzallah Azim Burney - 21701829 - CS
- Abdul Hamid Dabboussi - 21701076 - CS
- Mohamad Fakhouri - 21701546 - CS
- Sayed Abdullah Qutb - 21701024 - CS
- Hassan Raza - 21701811 - CS

**Supervisor:** Ercument Çiçek

**Innovation Expert:** Mustafa Sakalsız

# 1.0 Introduction

## 1.1 Overview

Social media platforms such as Twitter, Facebook, and YouTube are seeing a rise in the spread of fake news and misinformation on their platforms. According to research, Facebook users engage with misinformation 70 million times per month on average [1]. The rise of fake news is problematic for societies. It has the potential to sway public opinion, promote conspiracy theories, and instill fear, thereby eroding confidence in public institutions, and in democracy.

Therefore, our project aims to combat the spread of political lies and misinformation in order to better inform the public about what politicians are saying. To this end, we will build either a YouTube wrapper website that would perform fake news labelling and fact checking on live-streamed political speeches. The main targets of this project are political speeches and presidential debates from the United States of America. In order to accomplish this task, multiple ML/NLP papers on fake news will be consulted. A fake news/fact checking classification model will be built and trained on an appropriate dataset. Once a good enough classifier is achieved based on the metrics deemed ideal for such a problem, a platform would be built to use this classifier on political speeches streamed on YouTube. The model would then inform the viewer of any wrong claims being said by the speaker through various tags and captions.

## 1.2 Purpose of the System

## 1.3 Design Goals

The system's design goals are defined as follows,

### 1.3.1 Accessibility

- The website and its constituent services should be accessible by anyone on the internet. The website will be in the English language.

### 1.3.2 Usability

- The website should have a simple and intuitive interface with a minimalistic design.
- The website should be stable without interruptions and responsive enough to support *near* real-time feedback (5 minutes).

### 1.3.3 Compatibility

- The website should be compatible with different browsers and should work on both mobile and computer.

### 1.3.4 Scalability

- The website should be able to scale and handle upwards of 1000 users at a given time.
- The backend should be scalable in the sense that it should be able to run prediction models on 50 videos at a given time.

### 1.3.5 Performance

- Performance should be good and results should be obtained within 5 minutes.
- Overall response time of a website must also be instantaneous (less than 1 second).

### 1.3.6 Extensibility

- The website should be extensible so that in future it could accept videos from sources other than YouTube.

### 1.3.7 Accuracy

- The machine learning classifier’s “accuracy” would be measured using Macro-F1, Micro-F1 and Accuracy metrics to ensure the reliability of the model.

### 1.3.8 Reliability

- The website should be reliable to users and by being available all the time especially during live-streamed presidential speeches where traffic would increase substantially.

## 1.4 Definitions, Acronyms, and Abbreviations

| Term | Definitions, acronyms, and abbreviations |
|---|---|
| OAuth 2.0 | Google Sign-in and authentication |
| 3-Tier Architecture | ● Interface Layer (User interactions)<br>● Application Layer (ML Models)<br>● Data Layer (Database of facts) |
| noSQL | Non-relational database management for data storage retrieval |
| NLP | Natural Language Processing (Machine Learning Model) |
| FKIE | Fake News detection in social media posts such as Twitter. Created by Fraunhofer institute for Communication. [2] |
| Fabula AI | Fake News detection over text using a method called “Geometric Deep Learning”. Developed by a twitter-backed company based in London, called Fabula [3] |

# 2.0 Current Software Architecture

## 2.1 FKIE

FKIE is a fake news detection tool developed by Fraunhofer Institute for Communication, Information Processing and Ergonomics (FKIE), based in Germany. This tool checks social media posts, specifically twitter tweets for fake news. It processes texts and analyses the metadata of the tweet and then shows the findings in a visual form. The researchers at the Fraunhofer Society have developed a system that automatically analyzes social media sites and filters out false news and misleading information. This tool currently works on social media websites such as Facebook, and Twitter. [4]

The way researchers detect fake news using their build model can be categorized in 2 parts. The content, and Metadata. They used Buzzfeed Data training to train their models for the content of a social media post. BuzzFeed training is when a specific set of rules and examples are given to an algorithm and then add data to the dataset based on those rules and examples, and training starts that way. [5] This model of training is comparatively better because not all text posts need to be part of the dataset, which sometimes overfits the model. Training the model was done by BERT (Bidirectional Encoder Representations from Transformers) as a neural network architecture for sequence classification, in which the content of the social media post was processed in an unsupervised text dataset first and then it is trained on a labeled dataset and compared to a set of libraries which the researchers had prepared before. BERT is an attention-based architecture model for NLP (Natural Language Processing) [6]. These libraries were usually websites which mimicked original authentic news websites but with a difference in the wording of the news post itself. Different markers were prepared to use on training the dataset; These markers include punctuation mistakes, grammatical mistakes, spelling of the words, rarity of a word being used in a specific context or sometimes out-of-place expressions. This usually is a result of a non-native author writing a news article or social media post which links to a news article. [5]

**Metadata:**

Another marker used in this software is Metadata, which indeed plays a crucial role in differentiating between authentic sources of information and fake news. Metadata markers are good indicators of how accurate and true an article is. Metadata markers include the timing of a post, frequency of posts being issued, and geographical time zone of the author of the post. These markers all help in identifying whether a post is fake or not. For instance, if a user tweets about a topic in a really high frequency, such as 5 or 6 in a day, it increases the chance of being marked as fake news by the software.  Another metadata marker is the use of hate speech and racially discriminative words. [6]

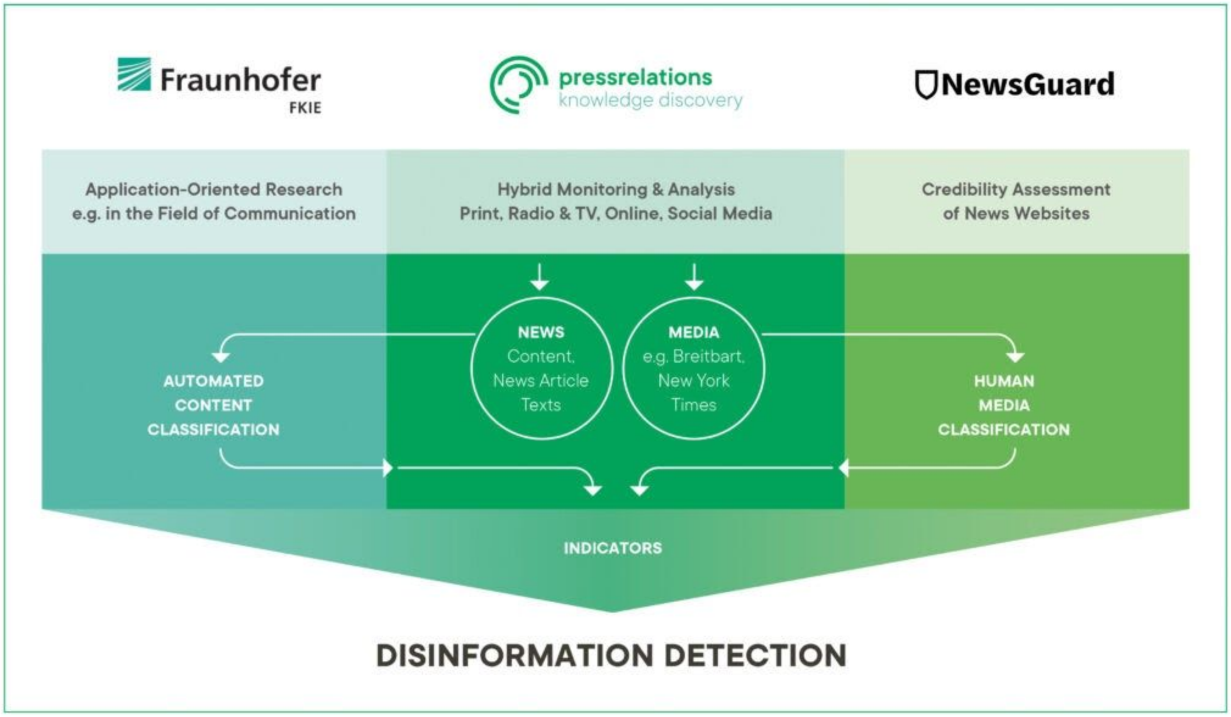
Fraunhofer FKIE
pressrelations knowledge discovery
NewsGuard
Application-Oriented Research e.g. in the Field of Communication
Hybrid Monitoring & Analysis Print, Radio & TV, Online, Social Media
Credibility Assessment of News Websites
NEWS Content, News Article Texts
MEDIA e.g. Breitbart, New York Times
AUTOMATED CONTENT CLASSIFICATION
HUMAN MEDIA CLASSIFICATION
INDICATORS
DISINFORMATION DETECTION

***Fig. 1:** The process of detecting disinformation [7]*

## 2.2 Fabula AI

Fabula AI is a twitter-backed startup company which developed a model to detect fake news in social media posts, specifically in Twitter tweets. [3] It was started as a startup in the United Kingdom in April 2018. Since then they have developed and patented a special model for training data on social networks. This model is called “Geometric Deep Learning” and is able to retrieve, and train data from any social media post in any language [8].

Geometric Deep Learning is a technique used for working on Non-Euclidean domains such as graphs and manifolds. Fabula AI developed this technique to analyze large datasets and figure out relations and interactions [9]. Fabula’s method of detecting fake news is different from FKIE’s product. The difference is that Fabula does not use NLP (Natural Language Processing) models to analyze the text of the post. It rather sees how a specific post is shared and interacted all over the social media, which requires the use of graphs, nodes and the connections, hence Geometric Deep Learning [10]. It analyzes datasets to find a pattern to how  news, both fake and real, is spread in the social media, which can be compared to how diseases spread on a network. [11]

Fabula AI is still in production stage and has not been released yet but aims to release an API later this year. However in the production stage, it trained on a dataset of 250,000 twitter users and 2.5 million tweets and the authenticity of each post was searched in Politifact and Snopes databases [10]. Currently Fabula AI claims that their model can achieve an accuracy of 93% within 2-20 hours of a post publishing [10].

# 3.0 Proposed System

## 3.1 Overview

This section provides a detailed overview of the different aspects of the project starting with subsystem decomposition to describe how different software components are arranged and interact with one another. Hardware/software mapping subsection describes where the different software parts of the project will live. Persistent data management subsection then describes how the project gets, handles and stores the data. How security is handled and how access control is granted is then explained in the access control and security subsection followed by a description of how the overall software is managed and controlled and how edge cases are handled.

## 3.2 Subsystem Decomposition

The system is designed following the 3 Tier Architecture. The Interface Layer contains all the boundary objects that the user can interact with and send all user provided input to the Application Logic Layer. The Application Logic Layer is responsible for transcribing the video, extracting relevant sentences, running sentences through the classifier and generating results. Results with high confidence scores are sent back to the Interface while results with low confidence are sent against a database query system which in turn sends back a result that is sent to the Interface. The Data Layer contains the Database Management component which checks low-confidence results against multiple fact-checkers to try and increase the confidence level and returns it’s result to the Application Logic Layer.

Below is the diagram representing this architecture.

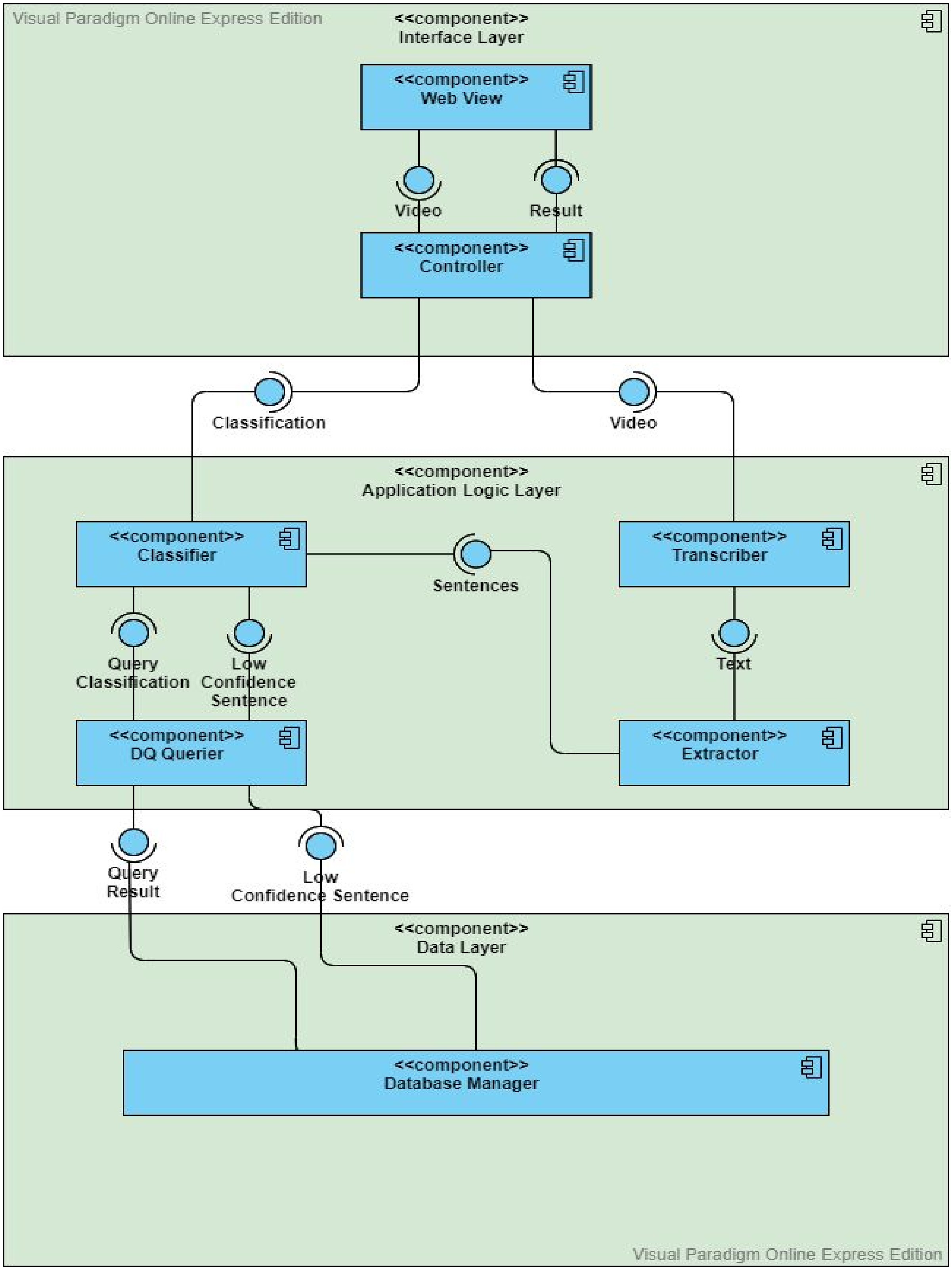

***Fig. 2**: 3-Tier Architecture*

## 3.3 Hardware/Software Mapping

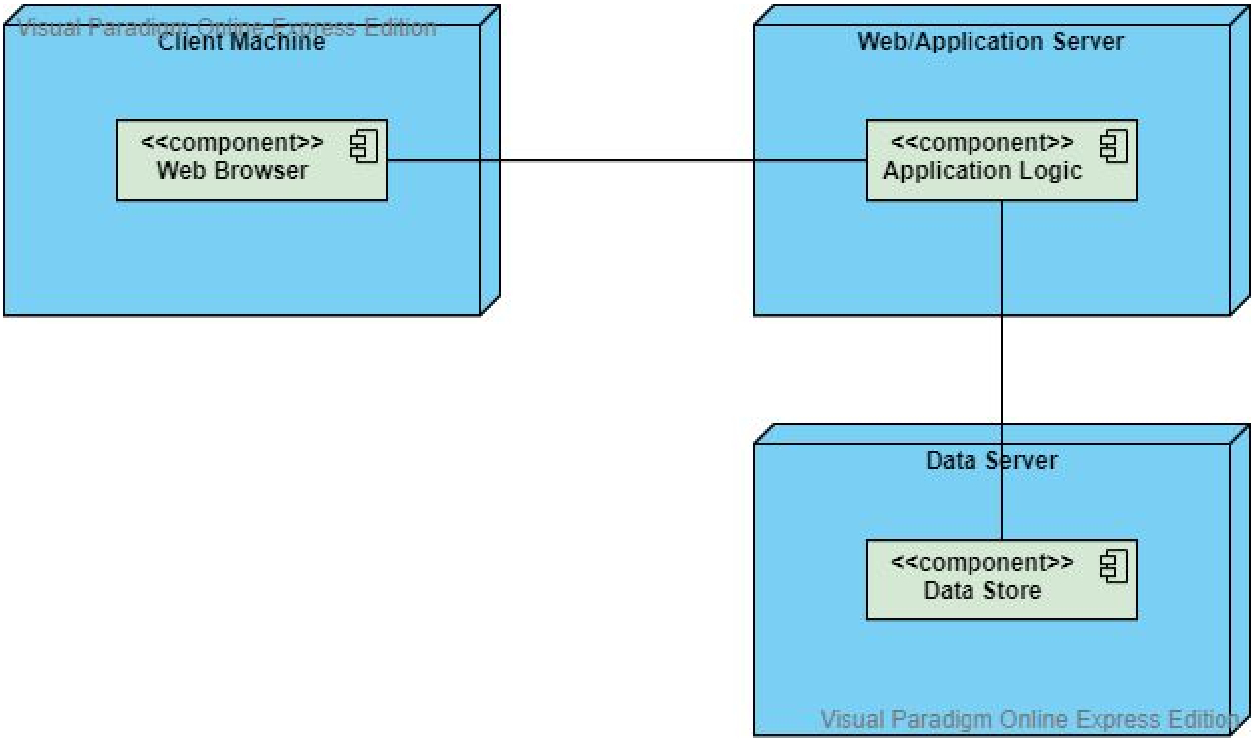

***Fig. 3**: Hardware/Software Mapping diagram*

The different system components discussed in the Subsystem Decomposition section will live in different places and on different types of hardware. All user interface components which are associated with a web browser will be on the client's machine while the application logic layer and the data layer will both live in the cloud on different servers (serverless technology will be used). The cloud part will be managed by the cloud provider and all communication between components will be managed using the providers fully managed services to insure scalability, elasticity, high availability and resilience.

## 3.4 Persistent Data Management

The Veritas system requires some data to be persisted over more than one single execution. The data that will be persisted includes saved facts, user account preferences, and fact reports generated for a video. The persistent data will be saved in a secure cloud noSQL database provided by Google Firebase. The document based model will provide a fast, efficient, and robust way of inserting, manipulating, and retrieving data to and from the database. It will also ensure that data is backed up in case of corruption or loss of any data on the database.

## 3.5 Access Control and Security

A user needs to have a Google account in order to create a profile in the system and access the necessary features. A user logs in to the web application using the aforementioned Google account credentials. The credentials are verified by the Google API and therefore, are not in any way stored or operated on by the system itself. After having gained access to their Veritas accounts, users can update their profile info, profile picture, generate fact check reports, stream political videos and get live fact checking on them, and be able to save facts that they might want to. In terms of data privacy, users cannot access any data of any other user, and by the design and premise of the system, data separability is guaranteed. The system doesn’t keep track of the videos watched by a user, or use their personal account information for any purposes. In order to bolster account security, the user will be given an authentication token when they login which will expire after 30 minutes. The user will then have to login again in order to access the system.

## 3.6 Global Software Control

Veritas makes use of event-driven software control. The system consists of two major actors namely, the client and the server. The system allows the client side to send certain requests to the server where they are processed, and the response is sent back to the client. In the request phase, the user enters the link to a YouTube video of a political speech and sends it to the server. The server receives the link, retrieves the video, and transcribes the video content using various data processing and NLP techniques in order to find the most relevant data needed to be classified via the machine learning model. The resulting text corpus is passed through the machine learning classifier and the resulting classifications are embedded through tags and caption back in the video. The video with the classifications are sent back as a response to the client where the video is rendered and the veracity of claims made in the speech are shown to the user. The user can also trigger an event to generate a report, and also call API methods to query specific sentences from other sources.

## 3.7 Boundary Conditions

### 3.7.1 Initialization

A user must have access, and be using the Veritas web application. A user must be able to sign-in using a Google account.

### 3.7.2 Termination

A user can log out of the Veritas web application or close the browser tab/window in order to terminate the system.

### 3.7.3 Failure

Users may experience several failures while using the Veritas web application. These may range from internet connection issues, video load failure, login failure etc. Our aim is to provide users with clearly understandable, informative, and transparent details about such errors, and to ensure the robustness of our application in dealing with these in a user-friendly manner.

# 4.0 Subsystem Services

## 4.1 Interface Layer

### 4.1.1 Web View

This component acts as the main page of the website. It is responsible for connecting the other components on the web page such as Video View, Results View, and Profile View.

### 4.1.2 Video View

A view to embed the video that is to be analyzed. It is responsible for displaying the video to the user with the controls.

### 4.1.3 Results View

A view that displays the sentences and their classification results. This is a scrollable list-view that contains two columns.

### 4.1.4 Profile View

This view allows the user to manage the profile settings.

### 4.1.5 Controller

The controller manages the main communication between the Web View and the Application Logic Layer. Its most important task is to send the video that is to be classified to the Logic Layer, and to receive the result and update the Results View. It also manages some interactions between the Views in the Interface Layer.

## 4.2 Application Logic Layer

### 4.2.1 Transcriber

The transcriber receives the video from the Controller in the Interface Layer and processes it to create text. This text will not be formatted in any way - just a collection of words. The transcriber then presents this text to the Extractor in the Application Logic Layer.

### 4.2.2 Extractor

The extractor takes the unformatted text from the Transcriber and cuts it into sentences. The exact method of how the text will be segmented has not been decided yet. This is a SBD (sentence boundary disambiguation) problem and one possible solution is to use DeepSegment [12]. Segmenting the text is important as the model predicts on claims (or sentences), not single words. The sentences then are passed to the Classifier.

### 4.2.3 Classifier

The classifier takes in the sentences from the Extractor and attempts to classify their truth value. We want to use a deep model for the classifier (an RNN) after we train it on the Multi-FC dataset. We will set up the model in such a way that it reports the confidence score for each result. If the confidence score is low, the classifier sends the sentence to the DB Querier. The DB Querier then returns the Query classification result and sends it to the Controller in the Interface Layer.

### 4.2.4 DB Querier

The DB Querier is responsible for querying the database manager for the claims that the model fails to predict accurately. The DB Querier then receives the query result, processes it and passes it to the Classifier back.

## 4.3 Database Layer

### 4.3.1 Database Manager

The Database Manager receives a sentence from the DB Querier and searches for it online by consulting a set of pre-specified sources. The result then is compiled and sent back to the DB Querier.

# 5.0 Consideration of Various Factors in Engineering Design

- **Public Health:** Veritas does not directly affect the health of its users.
- **Public Safety:** Veritas does not directly affect the health of its users.
- **Public Welfare:** The US presidential elections affect not only the United States, but the entire world. Hence, people from all countries follow the elections. Veritas is able to fact-check the claims that are being made in pseudo-real time, which in

turn allows the public to base their judgements on facts, not claims. This obviously increases the people's confidence in their judgements, which in turn increases the public welfare.

- **Cultural Factors:** Many topics that are popular in presidential debates are related to cultural factors. A recent article shows that few of the most popular debate topics are inequality, gender, education and gun laws [14]. All of these articles are of cultural importance to the public. The debaters usually state many claims regarding these  topics, and unsuspecting viewers might accept these claims without fact checking. By fact checking, Veritas provides the users with the ground truth, which helps in decreasing the false claims which are to be made in discussion of these topics.
- **Social Factors:** As discussed earlier, Veritas is capable of decreasing the amount of fake news and fake claims that spread as a result of the presidential elections. A recent study shows that fake news can lead people to create false memories especially if these fake news align with their beliefs [13]. This can strengthen prejudices in the society. By fact checking, Vertas can help in decreasing these false memories, in hope to improve the society as a whole.
- **Economic:** According to an article by Kenneth Rapoza, fake news can affect the stock markets heavily [14]. Veritas can decrease the amount of fake news, which can positively affect the stock market.
- **Environmental**: Veritas is served as a web application, and hence has no direct effect on the environment.

# 6.0 Teamwork Details

## 6.1 Contributing and functioning effectively on the team

We work as a team and all the responsibilities regarding different assignments and the actual making of application are divided among all the team members equally. Everyone tries to give their full input and work as hard as possible for doing the tasks. We have frequent team meetings in which everyone gives ideas and we have discussions about our project. Inputs are taken from all the team members when deciding about a certain thing and after that a decision is made with which all of us agree. Everyone is given the task in which he feels comfortable, or would like to take that task for the purpose of learning. Until now everyone has completed their tasks in time without any problems.

## 6.2 Helping creating a collaborative and inclusive environment

We have frequent meetings which all of us attend through online means or when possible in person. In these we discuss any new ideas one have or any new problems that one is facing. The problem is discussed and other team members help him in solving that problem. We discuss before taking any decision and take the opinion of all the team members so that any one member may not feel left out or offended that his opinion was not taken. At the beginning of semester we also went outside on a dinner together so that we could know each other better and can become familiar with each other, as some of us did not knew others previously.

## 6.3 Taking lead role and sharing leadership on the team

All the team members are responsible enough and know their responsibilities and try to complete them. However, a team must have a leader who can guide others and take the team together through the meeting and for completion of different tasks. We decided to share this responsibility, so each of our team members is made the leader of the team for two weeks, in these weeks he is responsible for arranging the meetings and giving tasks to different team members keeping in mind different deadlines we have so that we can finish our tasks on time. In this time as well other team members also help and remind the team leader about different things that he might be forgetting. But the decision about different things, if there is a dispute or miscommunication, is made by the leader. Leader is important inorder to avoid miscommunication.

# 7.0 Glossary

ML - Machine Learning

NLP - Natural Language Processing

API - Application Program Interface

f1 score - Harmonic mean of precision and recall

Macro f1 - Simple average of the f1 scores of each label

Micro f1 - Calculating precision and recall by summing all the True positives and Type errors instead of calculating for each label, and then calculating f1 score using these values.

# 8.0 References

[1] "Fake News, Misinformation, & Fact-Checking | Ohio University MPA | Ohio University", Ohio University, 2020. [Online]. Available: https://onlinemasters.ohio.edu/masters-public-administration/guide-to-misinformation-and-fact-checking/. [Accessed: 07- Oct- 2020].

[2] Admin, “A new software can detect fake news,” *Mathesia*, 18-Feb-1970. [Online]. Available: https://mathesia.com/a-new-software-can-detect-fake-news/. [Accessed: 27-Dec-2020].

[3] N. Lomas, “Fabula AI is using social spread to spot 'fake news',” *TechCrunch*, 06-Feb-2019. [Online]. Available: https://techcrunch.com/2019/02/06/fabula-ai-is-using-social-spread-to-spot-fake-news/. [Accessed: 27-Dec-2020].

[4] “Software that can automatically detect fake news,” *Fraunhofer*, 11-Feb-2019. [Online]. Available: https://www.fraunhofer.de/en/press/research-news/2019/february/software-that-can-automatically-detect-fake-news.html. [Accessed: 27-Dec-2020].

[5] T. S. World, “Software and Machine Learning Algorithms for Automatic Detection of Fake News,” *The Scientific World - Let's have a moment of science*, 01-Feb-2019. [Online]. Available: https://www.scientificworldinfo.com/2019/02/software-and-machine-learning-algorithms-to-detect-fake-news.html. [Accessed: 27-Dec-2020].

[6] A. Pritzkau, S. Winandy, and T. Krumbiegel, “Finding a line between trusted and untrusted information on tweets through sequence classification,” Unpublished, 2020, doi: 10.13140/RG.2.2.10902.37446.

[7] “pressrelations, Fraunhofer FKIE, and NewsGuard Join Forces to Detect Disinformation,” *NewsGuard*, 07-Jul-2020. [Online]. Available: https://www.newsguardtech.com/press/pressrelations_fraunhofer_fkie_und_newsguard_partnership/. [Accessed: 27-Dec-2020].

[8] M. Carella, “10 European startups protecting us from fake news and data privacy breaches,” *EU*, 09-Jun-2019. [Online]. Available: https://www.eu-startups.com/2018/11/10-european-startups-protecting-us-from-fake-news-and-data-privacy-breaches/. [Accessed: 27-Dec-2020].

[9] Flores Vivar, J. M. (2019). Artificial intelligence and journalism: diluting the impact of disinformation and fake news through bots. Doxa Comunicación, 29, pp. 197-212

[10] “NEWS CENTER,” *NVIDIA Developer News Center*, 11-Feb-2019. [Online]. Available: https://news.developer.nvidia.com/fabula-ai-develops-a-new-algorithm-to-stop-fake-news/. [Accessed: 27-Dec-2020].

[11] “Twitter snaps up Fabula.AI to tackle fake news: Articles: Big Data,” *Articles | Big Data | Innovation Enterprise*, 05-Jun-2019. [Online]. Available: https://channels.theinnovationenterprise.com/articles/twitter-snaps-up-fabula-ai-to-tackle-fake-news. [Accessed: 27-Dec-2020].

[12] “DeepSegment: A sentence segmenter that actually works” *Github*, [Online]. Available: https://github.com/notAI-tech/deepsegment [Accessed: 27-Dec-2020]

[13] “Fake News Can Lead to False Memories", Association for Psychological Science - APS, 2020. [Online]. Available: https://www.psychologicalscience.org/news/releases/fake-news-can-lead-to-false-memories.html. [Accessed: 21- Nov- 2020].

[14] K. Rapoza, "Can 'Fake News' Impact The Stock Market?", Forbes, 2020. [Online]. Available: https://www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market/?sh=1184c1612fac. [Accessed: 21- Nov- 2020].