GEORGE MASON UNIVERSITY

Data Analytics Engineering

Spring 2023

Advanced Data Tagging for Data Fitness and Quality

DAEN 690

Project Report

This page intentionally left blank.

Contents

Table of Contents

No table of figures entries found.

This page intentionally left blank.

Abstract

Abstract

Along with the explosive growth of data, high-quality data is required for analyzing and utilizing big data. Data quality and relevance are rarely combined when evaluating data fitness. In addition, business questions and tasks are not considered when determining data validity and authorization. This leads to high data management costs, poor decision making, additional resource requirements and compliance issues. This project seeks to leverage tags drawn from texts that allow the organization to automate the process of organizing their inventory of data assets with standard and structured tags and indicate fitness according to business tasks and questions.

The challenging part of the research area is achieving tagging with fitness involving quality and relevance. The proposed solution involves administrative tagging, domain tagging and topic modeling. Administrative tags describe data such as published date, person, organization, and location. In contrast, domain tags describe what topics best describe the content of the data. We utilized 5 different texts and image datasets collected from different sources. In case of text datasets web scraping was done to extract administrative tags, Latent Dirichlet Allocation (LDA) algorithm was applied by creating an ML/AI model to extract domain keywords/tags. On the other hand, for Image datasets, CNN model, confusion matrix, precision, and relevance applied to answer our business question. With the proposed models, we aim to find the covid related keywords from business questions from text dataset and to find which image dataset predicts covid information along with assessing the data quality and fitness indicators data.

This page intentionally left blank.

Report

Section 1: Problem Definition

1.1 Background

Big data is being accumulated at an unprecedented rate. It is too voluminous, fast, and complex to be processed using traditional methods. It comes from a multitude of origins such as consumer databases, transaction processing systems, healthcare data, social media, mobile applications, clickstream logs, etc. [1]. A common real-life example is Uber, as it creates and employs data about the drivers, vehicles, locations, trip distance, cost, and so on [2]. This data is explored to foresee the supply and demand situations.

1.1.1 Need for Data Management

While data comes in a well-structured form that can be stored in tables, big data is a totally different category. It does not always fit with well-defined labels [3]. The major challenges with Big Data are storage, processing, security, and data quality issues. We have too much data, but very little information [4]. Organizations have a hard time figuring out what they have in hand and when best to use it. This shows the need for data management.

Data management [5] is the procedure of gathering, sorting, securing, and storing an organization's data in a way to be scrutinized for business decisions. It is the pivotal step to carry out productive data analysis. Data management solutions are necessary to understand huge quantities of data. Data management enables people to quickly locate the information needed for analysis. It saves time and makes essential functions easier to perform.

1.1.2 Data Management Techniques

With effective data management, people can find the required data for making business decisions. It expands the visibility of the organization's data assets, thus making the company more organized and enabling employees to find the data they need. With reliable and the latest data, companies can adjust faster to market changes and customer needs. Data management also ensures protection against data losses, thefts, and breaches. It also lets organizations scale data and reduce repeated processes. The following are the different data management techniques [5]:

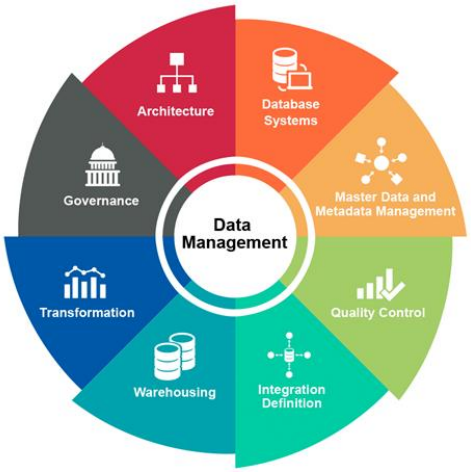

Figure 1: Data Management Techniques

- **Data Preparation**: Refers to cleaning and formatting raw data. Merging two or more datasets
- **Data Pipelines**: Ensures the automatic shift of data between different systems.
- **ETL** (Extract, Transform, Load): Formed to pull data from a system, transform it, and load it to the organization's warehouse.
- **Data Catalogs**: Assists in metadata management by laying out all aspects of the data including changes and locations, thus making the data easy to locate.
- **Data warehouses**: Integration of data origins happens here. It provides a comprehensible route for data analysis.
- **Data governance**: It specifies standards, procedure, and strategies to conserve data security and integrity.
- **Data architecture**: Provides a standard approach for creating and managing data flow.
- **Data security**: Protects data from uncertified access and attack.
- **Data modeling**: Registers the flow of data through an application or organization.

1.1.3 Metadata

Metadata [6] is the information that creates context for the data at hand. It makes it easier to organize and locate data to support business purposes. Metadata is created anytime a document, a file or other information is modified, including its deletion. For instance, consider a document. Its metadata elements are author, date created, date modified, and file size. It is important to note that metadata is just as crucial as the data itself. Precise metadata can be helpful in extending the lifespan of existing data by enabling users to discover new ways to apply it. There are three main types of metadata [7]: descriptive, administrative, and structural.

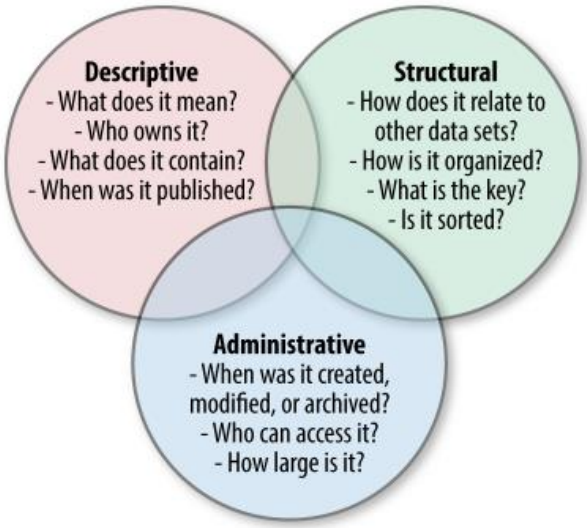

Figure 2: Types of Metadata

Descriptive metadata allows locating, recognizing, and selection of data assets. Components such as title, author, and subjects are included here. Administrative metadata eases the handling of resources. Elements such as technical, preservation, rights, and use are included. Structural metadata, commonly used in machine processing, narrates relationships among different parts of a resource.

1.1.4 Role of metadata in data management

Metadata plays a major role in data management. It makes sure that data is Findable, Accessible, Interoperable and Re-usable. (FAIR) [8]

By “Findable”, metadata makes it much easier to find related data. Most searches are done using text, so formats like audio, images, and video are restricted unless textual metadata are available. Text documents are also easier to find as the metadata explains exactly what the document is about. Data is “Accessible” because once someone finds the data they require, metadata shows how they can be accessed by including verification and authorization. By “Interoperable”, allocating metadata qualifies a data set to be unified with other data. They also make it easier for data to interface with applications or workflows for analysis, storage, and processing. To “re-use” a data set, people need to know the structure of the dataset, definitions of terms used, how they were gathered, and how they should be read or used. Data needs to be explained well, so that they can be duplicated and/or combined in different scenarios.

1.1.5 Data Discovery

While content search has always been around, it is just not exactly data discovery. Data discovery [9] involves the assembly and assessment of data from different roots and is used to understand drifts and patterns in the data. The intent of data discovery is to disclose appropriate data perceptions and communicate these insights in a comprehensible way, and eventually improve business processes.

Data discovery is a 5-step process [10], and it is iterative too. Organizations can carry on with collecting, analyzing, and refining their data discovery approach over time. The first step is to identify needs by beginning with a clear intent. Next, it is important to combine relevant data from different streams to get the whole story. This process is data crunching. Following this, the data needs to be cleansed and prepared, which is crucial. Noise is reduced in this step, and organizations get a transparent direction from the analyses. Step 4 is analyzing

the data, which has been merged from multiple departments and cleansed for analysis. The final step is to record the results and iterate when necessary.

1.1.6 Data Governance

Data Governance [11] is an important aspect that is responsible for ensuring that data is secure, private, accurate, accessible, and usable. It encompasses the actions that people must take, the processes that they must adhere to, and the technology that supports them throughout the data life cycle. Data Lifecycle Management (DLM) [12] extends to consider the end-to-end processing of data from ingestion through storage to transformation, analysis, and deletion. It contributes to the maintenance of data quality throughout its lifecycle, which enables process improvement and boosts productivity.

1.1.7 Data Cataloging, Data Tagging and Data Quality

It is a critical part of DLM that enables discovery and management of data resources. Additionally, it enables differentiation between similar data resources to indicate which data resource is more “authoritative” or a better fit. Our client, MITRE CORPORATION, a company, wants to develop data that is efficient and reliable for users. Due to the problems of manually tagging data, data is inconsistent and wastes resources. For data tagging to work efficiently, it needs to be adapted to auto-tagging [13]. Using ML/AI techniques to understand each dataset and automatically validate data quality while identifying ways to continually improve the quality of all datasets. So, we go on to explore how data tagging works. Data tagging [14] is the process of assigning tags or keywords to the various types of raw data (in the form of video, image, text, or audio) to make it simpler for the machine learning model to learn from. The Data Life Cycle is the sequence of data stages starting from its occurrence, how it is stored, used, shared, stored, and destroyed. Data tagging can improve the quality data life cycle. By data quality, we mean the capability of a dataset to be of use for the needs of an organization. Poor data quality leads to flawed analysis and lost revenue. Manual data entry errors, lack of complete information, ambiguous data, duplicate data, data transformation errors are the common causes of data quality problems. Data fitness [15] is the measure of how fit the data is to the problem at hand. It is described in 6 dimensions - Completeness, accuracy, timeliness, uniqueness, validity, and integrity.

1.2 Problem Space

Typically, companies do not consider and track quality and relevance in relation to business questions or tasks together while assessing data fitness. Data quality assessment is a systematic, business-driven approach to measuring and evaluating data quality dimensions, ensuring fitness for purpose, and establishing quality targets and thresholds to address specific business needs. Therefore, combining quality and relevance plays a very crucial role in assessing data fitness. Ideally companies can improve their ability for decision making, prediction based on ML/AI model when they combine quality and relevance for assessing data fitness.

Customer service, employee productivity, and important goals can all be seriously impacted when data collection fails short of the requirements set by the firm for correctness, validity, completeness, and consistency. Making accurate, informed decisions requires high-quality data. While all data has some degree of "quality," the level of data quality is determined by several qualities and factors (high-quality versus low-quality). Furthermore, distinct stakeholders throughout the firm will probably put a higher value on different data quality attributes [16].

Relevance is important when examining data quality traits since there must be a valid reason for gathering the data in the first place. A company should think about whether they need this information or if they are gathering irrelevant information or wasting time and money by gathering unrelated data [17]. Data tagging standards aid in managing and identifying sensitive data so that access to it can be properly regulated. Improved

data discovery will make it simpler to get data when needed. Identify the extent of data preparation required for any new data sources.

Better decision-making benefits, the entire organization when a business uses high-quality data to make business choices. A company may more effectively target its desired audience, whether internal or external, because of high-quality data. If not, businesses employ a scattershot strategy for their marketing, which results in higher costs and fewer sales. Knowing the audience better will help a company to create interesting content for its members. The more an organization understands its target market, the more they can develop campaigns that attract their interest in your offerings [18].

1.3 Research

Metadata is information about a data asset that characterizes it or provides information about it to make it easier to identify, assess, and comprehend. Data is systematically described through metadata, which also makes it easier for analysts to identify or connect important information across an organization. By allowing for the correlation of both old and new data, metadata also increases the lifespan of stored data. To design a customer journey, new product development, or competition analysis effectively and precisely, it raises the value of the organization's information resources.

Most organizations are currently dealing with the challenges posed by the growing number of institutional digital archives and are attempting to determine whether metadata can provide solutions to these issues. As a result, metadata can provide a solution to enterprise needs by improving access to critical information contained in large documents. However, manually tagging data with metadata is a time consuming, resource consuming and costly function requiring specialized knowledge and training to ensure that metadata accurately realizes data discovery, access, and use. As the volume of labeled data increases, it is necessary to publish robust metadata with respect to data suitability and quality. To be data efficient, data tagging must be tuned to auto-tagging using machine learning and artificial intelligence techniques to understand each dataset and automatically validate data quality while identifying ways to continuously improve the quality of all datasets [19].

The quality of metadata describing digital resources stored in a repository can be regarded as a necessary condition for the repository's reliable and efficient operation. Metadata is widely regarded as the key to successfully locating relevant resources. As a result, metadata must be created and maintained in accordance with predefined procedures. This requirement is even more important given the vast number of available digital resources, which is rapidly increasing. Even though the need for quality metadata is widely acknowledged, there is no widely accepted approach to defining metadata quality and, as a result, how this quality can be assessed, measured, and improved [20].

According to research we conducted to understand the importance of data tagging and data quality monitoring. Many studies have attempted to categorize and systematize data to maintain data quality because decision-making and analysis require consistency, accuracy, completeness, uniqueness, and timeliness [21]. The goal of data quality management is to detect errors in the data and fix them. One of the most widely accepted is that raw data categorization distinguishes between single-source and multi-source data problems. The two types are further classified as schema-level and instance-level. Different schema and data model problems are involved in schema-level problems. In contrast, instance-level problems are caused by input errors [22].

1.3.1 Assessing Data Quality

A method for determining the suitability of a given data set for use by mapping the quality dimensions to specific data quality issues in terms of consistency, accuracy, completeness, uniqueness, and timeliness [21]

[22]. For example, missing value issues violate validity and integrity, while copy availability violates the uniqueness dimension [22]. This method evaluates the errors that occur in the data set to determine which dimension of quality problems. The factors that contribute to the error for each dimension are as follows (1) **Consistency**: ambiguous data, extraneous data, misfielded values, structural conflicts, different word orderings, different aggregation levels, different units' representations, functional dependencies violation, wrong data type, and referential integrity violation. (2) **Accuracy**: missing data, incorrect data, misspellings, ambiguous data, misfielded values, incorrect references, different aggregation levels, temporal mismatch, domain violation, functional dependencies violation, and referential integrity violation. (3) **Completeness**: missing data, misfielded values, and referential integrity violation. (4) **Uniqueness**: extraneous data, duplicates, structural conflicts, different word orderings, uniqueness violation, and use of synonyms. (5) **Timeliness**: outdated temporal data, and temporal mismatch.

We concluded that a particularly concerning issue that affects the quality of data in many dimensions is that misfielded values are values that are placed inside the wrong attribute and referential integrity violation is multiple-relationship tuples that violate the referential integrity constraints or missing foreign key.

1.3.2 Assessing Data Relevance

Search relevance is the measure of accuracy of the relationship between the search query and the search results. Search relevance is an assessment of how closely a search result is related to the query. High search relevance means users can easily find the right information at the right time. The degree to which search results are relevant and helpful to the user's search query is referred to as search relevance. In other words, it is a measure of how well the search engine comprehended the user's intent and delivered results that met their requirements and expectations. The two most basic metrics for measuring this are **precision** and **recall**.

- Recall describes the percentage of relevant documents within the corpus that are returned in the results, so we want to ensure all relevant documents of a dataset are included in the search results.

$$recall = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ relevant\ documents}\ X\ 100$$

- Precision describes the percentage of relevant documents within the returned results, so we want to ensure all data in the search results are relevant.

$$precision = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ documents\ retrieved}\ X\ 100$$

Assume you have 100 documents in total, 25 of which are pertinent to a query. Assuming you returned all 25 documents properly, but also 5 unrelated documents. The following is how precision and recall would be calculated: Precision: 25 retrieved relevant / 30 total retrieved = 0.8 and Recall: 25 retrieved relevant / 25 total relevant = 1

Another way to assess relevance is Term-Document and the Bag-of-Words Matrix to improve on our basic search model, we need a metric to compare search results with.The most obvious answer is to examine word frequency. If one document includes our search terms frequently and another document only a few times, the first is likely to be more relevant. The bag-of-words (BoW) can be used to examine word frequency. A BoW converts a text into a set of token-frequency tuples. This collection can be thought of as a vectorized version of

our text, which can be helpful in a variety of machine learning tasks. We can convert it into a term-document matrix, which is a table with each row representing a different word.

1.4 Solution Space

Our process helps users and organizations to incorporate data fitness measures or indicators into DLM processing that consider business questions and tasks by using meta data tags. Data tagging is assigning tags or keywords to the various types of raw data (video, image, text, or audio) to make it simpler for the machine learning model to learn. The Data Life Cycle is the sequence of data stages starting from its occurrence and how it is stored, used, shared, stored, and destroyed. Data tagging can improve the quality data life cycle. We use DLM processing to include data suitability metrics or indicators that consider business issues and tasks.

The variety of data sources results in a wide variety of data types and intricate data structures, which makes data integration more challenging. Data checks, also known as metrics, can be used to describe a dataset's characteristics, such as missing values, outliers, and frequency distributions.

Our process is shown in the business question in the schematic flow diagram presented in section 3.3.1. We will use Momentum as one of the platforms. From Business Questions and Data Resources, we map Business Questions DLM types and steps from there DLM processing steps start from Business Question to DLM maps it assesses the relevance and then generate fitness indicators from their mapping for authoritative it connects to business question to DLM maps. For details, refer to Section 3.1.1 Systems Architecture.

1.5 Project Objectives

The team, in consultation with MITRE, recognized and developed our objectives for the successful completion of the project, which are detailed below with supporting steps.

1.5.1 Business Objective

The primary objective of this project is to better enable users to differentiate datasets that are more likely to meet their business needs. This helps users and organizations to save time and resources to refocus human efforts toward tuning parameters instead of data entry of metadata.

1.5.2 Technical Objective

The technical objective is to provide a framework by which to link business questions and tasks to better differentiate data using advanced metadata tagging and indicators for data fitness; thereby, reducing manually intensive work. It will ensure the metadata accurately enables discovery and usage of the data for data fitness, which consists of measuring quality and relevance. This framework utilizes AI/ML techniques; albeit, in an agnostic manner to allow specific algorithms to be swapped in and out to better address a wide range of business needs.

We establish a Business Question Mapping of data resources characteristics and Data Lifecycle Management (DLM) steps (ingest, condition, discover, search, govern, prepare for analytics). Combining data quality and data relevance in metadata tagging will ensure robust business fit and performance. Additionally, we will generate tags, indicators (accessibility, usability, reliability, consistency, relevance, fitness, presentation quality), and measures appropriate to indicate data fitness. This will enable differentiation between similar data resources to indicate which data resource is more authoritative or a better fit.

This mechanism automates the tagging of data, where a different style of tagging is likely to indicate measurement or degree of fitness [or something like this] or what you say in next paragraph might suffice.

1.5.3 Learning Objective

In terms of understanding the problem, we aim to understand the data fitness and data quality for the metadata, different structures, and characteristics of the data. Furthermore, to create meta tags from structured, semi-structured, and unstructured data.

The learning objective of this project is to learn to identify the indicators of data quality and automate the process of creating the tags. We would be using advanced data tagging with metadata for data fitness and quality to reduce the manually intensive data management life cycle functions (ingest, condition, discover, search, govern, prepare for analytics). This will help refocus human efforts toward tuning parameters instead of data entry of metadata.

1.6 Primary User Stories

"Based on the user context and value proposition, we developed the following primary user story to guide our project.

"As a company, after collecting users' data, before using it, we want to differentiate datasets that best address our business questions and tasks." which can be achieved if tags and metadata are in place.

After collecting the data, we will assess the fitness and quality of the data. Using a data quality tag, the decision-maker is notified of the data's quality. We will use advanced data tagging with metadata for fitness and quality to reduce manual work operations.

1.7 Product Vision

It is anticipated that data quality and fitness play a major role in data analysis. This trend is almost all followed by every data dealing company. Since data is stored and managed in large scales it would be better to organize data in a simpler way and easy to use or refer to it. Our product helps to check the fitness of the data and generates metadata tags, and these tags can be used for further analysis according to requirement. This process includes data dealing with images and texts. This platform is being tested on multiple datasets.

For: Mapping business questions and assessing data quality and relevance

Who: Wants to have quality data to fit into their requirement

The: Image data, text data

That: Helps them to track business questions and tasks and assess data fitness.

Our product: Our solution will help users to easily differentiate datasets that best address their business questions and tasks and assess data resources for fitness.

Caveats: The processing and training takes a tremendous amount of computational power and time for accuracy of the meta data.

1.7.1 Scenario #1

An ed-tech company wants a text-based prediction of images. For the easy and clear understanding of the concepts taught by a professor to students. Our product helps to make sure the data fitness and quality standards are met for the advance data tagging for the metadata.

1.7.2 Scenario #2

An Encyclopedia company wants to gather data based on text annotations. Our product helps to keep the data in an organized manner with the help of advanced data tagging by making sure the data quality and fitness are met with the company requirements and further the data can be processed by training the dataset based on the texts using computer vision and ml.

Section 2: Datasets

2.1 Overview

This section briefly discusses the domain of each dataset and where we acquired it. Initially, we selected three datasets suitable for the data tags domain. These datasets are in semi structured, and unstructured formats. Our aim is to see how we can use different forms of data to automatically tag and assess the data fitness and quality. The below mentioned datasets could be used in various stages of our data modeling.

Table 1. Dataset Summary

Dataset	Owner	Type	Size	Format
MITRE Synthetic Health Data	MITRE Corporation	Open Source	9 GB	image
BBC News	Kaggle	Open source	1 MB	csv
Tweets	Data.world	Open Source	1 GB	text
Covid-19 Image	Kaggle	Open source	166 MB	Image
Covid CXR Image	Kaggle	Open source	600 MB	Image

MITRE Synthetic Health data is a comprehensive collection of realistic, artificial patient data. It was created by MITRE Corporation to help medical research and innovation. Since this is artificial patient data, there is zero risk of data leakage or privacy concerns. Each detail is created from a model that researchers can adjust to meet their specific needs.

BBC News dataset self-updates each week by including RSS feeds. This data is semi-structured. The data consists of the following columns - title, pubDate, guide, link, and description. It gathers RSS Feeds from the BBC News website using requests html and BeautifulSoup. The dataset for Transactions has 13613 records. We can utilize the data to examine the sentiment of the news.

Tweets dataset is a semi-structured dataset. This dataset contains millions of Tweets related to the Coronavirus/COVID-19 pandemic to find possible patterns of disinformation and unusual propagation of content on Twitter. It includes more than 19.95M tweets mentioning Coronavirus and/or COVID-19 using the free Twitter public API in separate time periods starting in early February 2020. Millions of tweets, blog posts, and news stories written in many languages are included in the Natural Language Processing dataset. The datasets for Tweet blogs contain 4 million texts, news, blogs totaling around 1 GB. The texts have a variety of sentences.

The COVID-19 Image Dataset is a collection of chest X-ray images from COVID-19 patients and healthy individuals who had viral pneumonia. A total of 317 images are in the group, of which 137 are COVID-19 positive, and the remaining 180 are either viral pneumonia images or standard chest X-rays. For machine learning, the images are organized into train and test directories.

The COVID CXR Image Dataset includes 1823 pictures of chest X-rays' annotated posteroanterior (PA) views. For viral pneumonia, non-pneumonia cases, and normal cases, labeled optical coherence tomography (OCT) and CXR images are used. The dataset includes 619 images of viral pneumonia, 536 images of COVID-19, and 668 typical patients.

2.2 Field Descriptions

2.2.1 MITRE SYNTHETIC HEALTH DATA

MITRE Synthetic Health Data was created in 2018 for the purpose of medical studies and innovation. Realistic artificial patient data containing attributes that could identify patient information is generated by the tool SYNTHEA. This ensures that there is no data leakage or ethical issues when working with sensitive medical information.

Field Name	Description	Data Type	Source
ID	This is the unique patient ID that is used to identify patient and patient information. This ID is initially generated when a patient is admitted, or the record is made.	String	Original
BIRTHDATE	This is the birth date of the patient. The format is in mm/dd/yyyy. For example: 09/29/1982. We do not have any null records or in correct formats for this field.	Date	Original
SSN	This is the social security number of the patient. It is a numerical identifier assigned to U.S. citizens and other residents to track income and determine benefits. It has 9 digits. For example: 999-32-7680. This field does not have null.	Number	Original

DRIVERS	This is the driver's license of the patient. It is a specific identification number assigned to a driver by the issuing government agency. It has 9 characters. For example: S99960247. This field has values for only the patients who have driver's licenses.	String	Original
PASSPORT	This is the passport number of the patient. It is a unique identification number, and it has null values for the patient who does not have a passport number.	String	Original
FIRST	This is the first name of the patient. This might not be unique as multiple patients can have the same first name. For example: Lisabeth, Anisaa.	String	Original
LAST	This is the last name of the patient. This might not be unique as patients can have the same last names. For example: Fadel, Kris.	String	Original
MARITAL	This is the marital status of the patient. It is 'M' for married, 'S' for Single. It has blank fields for patients whose marital status is unknown.	Character	Original
RACE	This is the patient's race. It could be Black, White, Asian, Native etc.	String	Original
ETHNICITY	This is the patient's ethnicity. This indicates if the patient is hispanic or non-hispanic. This field does not have null values.	String	Original
GENDER	This field indicates the patient's gender. This field either has 'M' for Male or 'F' for Female.	Character	Original

BIRTHPLACE	This field has the patient's birthplace. It includes the place and state in which the patient is born. For example: Belmont Massachusetts.	String	Original
ADDRESS	This field has the exact place where the patients are living which contains the community or apartment name with the door number. For example, it could be 383 Crook's cramp.	String	Original
CITY	This field could be any city where the patients are living in Massachusetts state. For example, it could be Boston or Lynn.	String	Original
STATE	This field is the state where the patients are living. For this Dataset all the entries are from Massachusetts state.	String	Original
COUNTY	This field is the County in Massachusetts state where the patients are living. For example, it could be Bristol County or Essex County.	String	Original
ZIP	This field is the ZIP code (postal code) for the place where the patients are living. For example, it could be 1702 or 2109.	String	Original
LAT	This field is the longitude of the place where the patients are living. For example, it could be 42.15189.	Float64	Original

LON	This field is the latitude of the place where the patients are living. For example, it could be -72.19878.	Float64	Original
HEALTHCARE_EXPENSES	This field contains the total expenses of the individual patient in dollars. For example, it could be 8466.29.	Float64	Original
HEALTHCARE_COVERAGE	This field contains the total health care provided for each patient in dollars. For example, it could be 1588.45	Float64	Original

2.2.2 BBC News

BBC news dataset created by Gabriel Preda was published on 2022-02-18. BBC News RSS Feeds is a list of news titles from Kaggle [23], which is updated daily, from 2022-02-18 until now with news links and descriptions. There are 5 different field names with each field name being unique as described. All data types are objective, and all come from original sources.

Field Name	Description	Data Type	Source
Title	This is a news title, which is unique and used to identify the news headlines and subheadings.	objective	Original
PubDate	This is the exact date and time each news was published, along with the date and time, with GMT time zones. For example, Mon, 07 Mar 2022 00:02:15 GMT.	objective	Original
Guid	This is a guid for each BBC news that lists links to the BBC News online site. For example, https://www.bbc.co.uk/news/technology-60608222	objective	Original
Link	This is a link to each BBC news that lists links to the BBC News online site. For example, https://www.bbc.co.uk/news/technology-60608222?at_medium=RSS&at_campaign=KARANGA	objective	Original

Description	This is a brief description of each news briefly summarizing the contents of each news in just 1-2 sentences.	objective	Original

2.2.3 Tweets

Tweets Blogs News is a zip file created by SWIFTKEY in collaboration with the Johns Hopkins Data Science Specialization from Kaggle [25]. Millions of tweets, blog posts, and news articles in multiple languages are included in this Natural Language Processing dataset.

Field Name	Description	Data Type	Source
Body	This column contains the quotes of any person who interacted in quotation marks.	String	Original
URL	This column is about news stories and situations.	String	Original

2.2.4 Covid-19 Image Dataset

It is a simple directory structure branched into test and train and further branched into the respective 3 classes which contains the images. It contains around 137 cleaned images of COVID-19 and 317 in total containing Viral Pneumonia and Normal Chest X-Rays structured into the test and train directories.

Field Name	Description	Data Type	Source
Covid	This folder has images related to X-rays of covid affected patients.	.dcm	Original
Normal	This folder has images related to X-rays of patients who have normal conditions.	.dcm	Original
Viral Pneumonia	This folder has images related to X-rays of patients infected with Viral Pneumonia.	.dcm	Original

2.2.5 Covid CRX Image Dataset

The COVID CXR Image Dataset is a collection of chest X-ray images that have been categorized into three classes namely Covid, Normal, Viral Pneumonia. There are a total of 1823 CXR images.

Field Name	Description	Data Type	Source
Covid	This folder has images related to X-rays of covid affected patients	.dcm	Original
Normal	This folder has images related to X-rays of patients who have normal conditions.	.dcm	Original
Virus	This folder has images related to X-rays of patients infected with Virus.	.dcm	Original

2.3 Data Context

MITRE Synthetic Health Data is a collection of 16 csv files that contains patient information specific to procedures, encounters, hospitalizations, immunizations, observations, equipment supplies and so on. From these files, various conditions could be analyzed, interpreted, and visualized.

The tweets dataset being analyzed here is highly unstructured. With each tweet being limited to 280 characters, Twitter has proved to be a ground for successful marketing strategies. Through Twitter data, important attributes such as profile visits, mentions, tweet impressions, tweet engagement rates, the top trending tweets, follower growth, video content performance and content tracking [26]. Using Twitter's built in analytics tool, several factors important for businesses can be studied – Account performance summary, audience growth, publishing behavior, audience demographics, statistics by profile, etc. [27]. Through this, businesses can get a picture of their core buyers and find out what marketing strategies work.

BBC news is the next dataset that is analyzed. It updates on its own by collecting live feeds using a kernel [23]. Data is collected using HTML and Python's BeautifulSoup. The overall sentiment of the news, title and description are gathered.

Covid-19 image dataset can be utilized for educational purposes, particularly for building machine learning algorithms for COVID-19 detection from chest X-rays [28]. It's crucial to remember that this dataset is relatively small and could not be entirely representative of the population. Additionally, it's important to remember that machine learning models shouldn't be used in place of qualified medical personnel when diagnosing or treating diseases.

2.4 Conditioning

Data conditioning is an essential phase in the pre-processing of a dataset, which is followed by data management and optimization techniques that result in intelligent routing, optimization, and protection of data for storage or data mobility within a system. Data conditioning incorporates outlier identification, missing/incomplete data, data substitution and noise removal. It is traditionally done using statistical methods, but a group of machine learning techniques is also used based on the data. There are a variety of methods for data conditioning that would facilitate the extraction of this product's metadata. The Data Pipeline stage is the initial phase. There are three text datasets and two image datasets. For both text and images, data cleansing is a crucial step.

2.4.1 MITRE Synthetic Health Data

We have selected six datasets from 100k MITRE datasets which are Observations, Patients, Procedure, Care Plans, and Encounters. We worked collaboratively with other two teams to merge those datasets together into one dataset. As we three teams collaboratively worked together, we mashed up datasets as per our business requirements.

2.4.2 BBC News

This dataset includes five columns, only three columns can be used to automatically tagging data. Based on the data, we notice that each row has duplicated data, it needs to process these duplicated rows, there are no missing values. We changed Pubdate format to YYYY-MM-DD format. Then dropped duplicate values from 13613 rows × 5 columns to 12324 rows × 5 columns.

```
In [8]: data = data.drop_duplicates(subset=['title']).drop_duplicates(subset=['guid']).drop_duplicates(subset=['description']).reset_inde
        data
```

Out[8]:

	title	pubDate	guid	link	description
0	Ukraine: Angry Zelensky vows to punish Russian...	2022-03-07 08:01:56	https://www.bbc.co.uk/news/world-europe-60638042	https://www.bbc.co.uk/news/world-europe-606380...	The Ukrainian president says the country will ...
1	War in Ukraine: Taking cover in a town under a...	2022-03-06 22:49:58	https://www.bbc.co.uk/news/world-europe-60641873	https://www.bbc.co.uk/news/world-europe-606418...	Jeremy Bowen was on the frontline in Irpin, as...
2	Ukraine war 'catastrophic for global food'	2022-03-07 00:14:42	https://www.bbc.co.uk/news/business-60623941	https://www.bbc.co.uk/news/business-60623941?a...	One of the world's biggest fertiliser firms sa...
3	Manchester Arena bombing: Saffie Roussos's par...	2022-03-07 00:05:40	https://www.bbc.co.uk/news/uk-60579079	https://www.bbc.co.uk/news/uk-60579079?at_medi...	The parents of the Manchester Arena bombing's ...
4	Ukraine conflict: Oil price soars to highest l...	2022-03-07 08:15:53	https://www.bbc.co.uk/news/business-60642786	https://www.bbc.co.uk/news/business-60642786?a...	Consumers are feeling the impact of higher ene...
...	...	...	...	...	...
12319	One ruined neighbourhood at the heart of devas...	2023-02-09 11:27:55	https://www.bbc.co.uk/news/world-europe-64581229	https://www.bbc.co.uk/news/world-europe-645812...	Thousands are sheltering in makeshift camps li...
12320	Turkey-Syria earthquake: Watching the search f...	2023-02-09 11:15:30	https://www.bbc.co.uk/news/uk-64581995	https://www.bbc.co.uk/news/uk-64581995?at_medi...	As rescuers look for earthquake survivors in T...
12321	Turkey-Syria earthquake: First aid convoy reac...	2023-02-09 16:27:47	https://www.bbc.co.uk/news/world-middle-east-6...	https://www.bbc.co.uk/news/world-middle-east-6...	Vital UN deliveries to the opposition-held reg...

Figure 3: Removing Duplicates

Additionally, we removed punctuation for the title. The title text has several punctuations. Punctuation is often unnecessary as it doesn't add value or meaning to the NLP model.

```
In [36]: def clean_text(text):
             text = str(text).lower()
             text = re.sub('\[.*?\]', '', text)
             text = re.sub('https?://\S+|www\.\S+', '', text)
             text = re.sub('<.*?>+', '', text)
             text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
             text = re.sub('\n', '', text)
             text = re.sub('\w*\d\w*', '', text)
             return text

In [38]: data['title'] = data['title'].apply(clean_text)
         data.head()
```

Out[38]:

	title	pubDate	guid	link	description
0	ukraine angry zelensky vows to punish russian ...	2022-03-07 08:01:56	https://www.bbc.co.uk/news/world-europe-60638042	https://www.bbc.co.uk/news/world-europe-606380...	The Ukrainian president says the country will ...
1	war in ukraine taking cover in a town under at...	2022-03-06 22:49:58	https://www.bbc.co.uk/news/world-europe-60641873	https://www.bbc.co.uk/news/world-europe-606418...	Jeremy Bowen was on the frontline in Irpin, as...
2	ukraine war catastrophic for global food	2022-03-07 00:14:42	https://www.bbc.co.uk/news/business-60623941	https://www.bbc.co.uk/news/business-60623941?a...	One of the world's biggest fertiliser firms sa...
3	manchester arena bombing saffie roussoss paren...	2022-03-07 00:05:40	https://www.bbc.co.uk/news/uk-60579079	https://www.bbc.co.uk/news/uk-60579079?at_medi...	The parents of the Manchester Arena bombing's ...

Figure 4: Tokenization

Next, we tokenized the 'text'. Tokenizing is the process of splitting strings into a list of words. After the tokenization we removed the stopwords. Stop words are irrelevant words that won't help in identifying a text as real or fake. We used the "NLTK" library for stop-words.

```
In [13]: def remove_stopwords(text):
             text=[word for word in text if word not in stopword]
             return text
         data['title_wo_punct_split_wo_stopwords'] = data['title_wo_punct_split'].apply(lambda x: remove_stopwords(x))
         data.head()
```

Out[13]:

	title	pubDate	guid	link	description	title_wo_punct_split	title_wo_punct_split_wo_stopwords
0	ukraine angry zelensky vows to punish russian ...	2022-03-07 08:01:56	https://www.bbc.co.uk/news/world-europe-60638042	https://www.bbc.co.uk/news/world-europe-606380...	The Ukrainian president says the country will ...	[ukraine, angry, zelensky, vows, to, punish, r...	[ukraine, angry, zelensky, vows, punish, russi...
1	war in ukraine taking cover in a town under at...	2022-03-06 22:49:58	https://www.bbc.co.uk/news/world-europe-60641873	https://www.bbc.co.uk/news/world-europe-606418...	Jeremy Bowen was on the frontline in Irpin, as...	[war, in, ukraine, taking, cover, in, a, town,...	[war, ukraine, taking, cover, town, attack]
2	ukraine war catastrophic for global food	2022-03-07 00:14:42	https://www.bbc.co.uk/news/business-60623941	https://www.bbc.co.uk/news/business-60623941?a...	One of the world's biggest fertiliser firms sa...	[ukraine, war, catastrophic, for, global, food]	[ukraine, war, catastrophic, global, food]
3	manchester arena bombing saffie roussos	2022-03-07 00:05:40	https://www.bbc.co.uk/news/uk-60579079	https://www.bbc.co.uk/news/uk-60579079?at_medi...	The parents of the Manchester Arena bombing's	[manchester, arena, bombing, saffie, roussos,...	[manchester, arena, bombing, saffie, roussos,...

Figure 5: New dataset (without punctuation, without stop words and split strings)

2.4.3 Covid-19 Image

The collection consists of 317 images, of which 137 are COVID-19 positive, and the other 180 are either viral pneumonia images or standard chest X-rays. For machine learning, the images are divided into train and test directories.

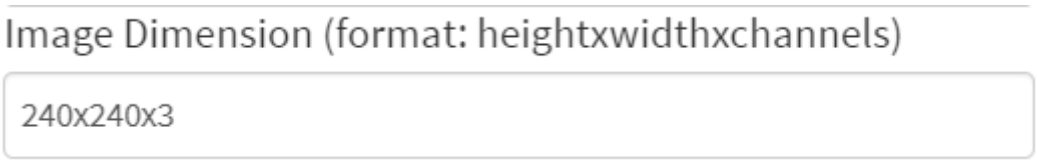

Figure 6: Resizing of Covid-19 Image Dataset

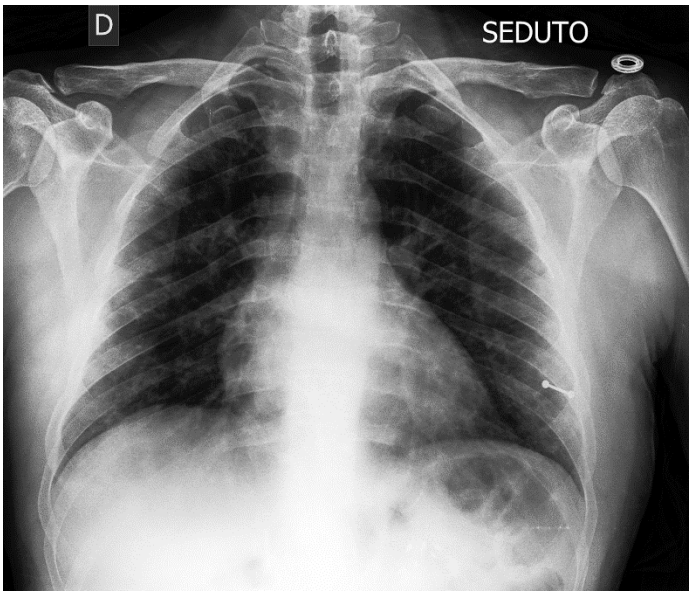

Figure 7: X-ray of Covid Patient from Covid-19 Image Dataset

2.4.4 Covid CRX Image

This dataset includes patients who are Normal, Viral, and COVID-19-affected in a posteroanterior (PA) view of chest X-ray pictures. There are 1823 CXR pictures in all. The dataset can be used for research purposes to create machine-learning models for COVID-19 detection from chest X-rays. It's crucial to remember that machine-learning models shouldn't be utilized in place of qualified medical personnel when it comes to diagnosing or treating illnesses.

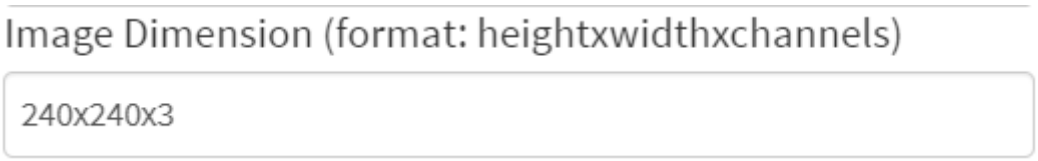

Figure 8: Resizing of Covid CXR Image Dataset

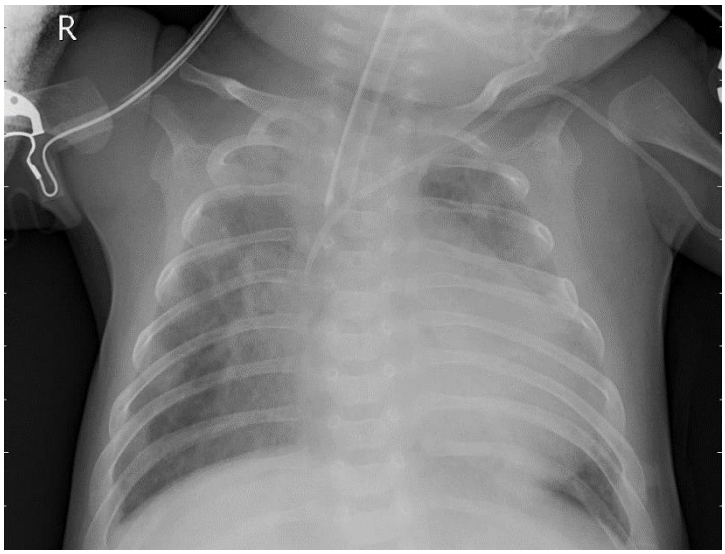

Figure 9: X-ray of Virus infected patient from CXR Image Dataset

2.5 Data Quality Assessment

Dataset 1: MITRE Synthetic Health Data

Completeness – This dataset has a medium degree of completeness since there is a significant quantity of missing data, such as prefix, death date, suffix, and maiden; nonetheless, nearly all the found have complete data for that personal information data.

Consistency – there are 25 columns with a static data type and consistent format for appropriate columns, such as SSN '999-68-6630.'

Uniqueness – This dataset has a high level of uniqueness since there are no duplicates.

Conformity – All of the fields are in consistent form.

Accuracy – This dataset has a high level of accuracy because it was created using SyntheaTM, an open-source patient population simulation made accessible by The MITRE Corporation, and it is generated by SyntheticMass.

Overall Quality – In general, given the missing data, the overall quality of this dataset is acceptable, although there are some issues we might be concerned about.

Dataset 2: BBC News

Completeness–This dataset has a high degree of completeness as there is no missing data. Dataset has 13613 records with 5 columns. Every column has 0% of null values, indicating that there are no missing values.

Consistency– This dataset is highly consistent. Because every column is an objective data type which is title, pubdate, guid, and link. There is no extraneous data and no reference integrity violations.

Uniqueness–The dataset is unique and has no duplicate values, this dataset has a high level of uniqueness.

Integrity–There is no entity integrity because there is no primary key that uniquely identifies a record such as Title _Id.

Conformity–Pubdate column does not fit conventions because dates should be datetime64 in YYYY-MM-DD format.

Accuracy– There are no outliers in the dataset and no misspellings which indicate this dataset has a high degree of accuracy.

Overall Quality – The overall dataset is a high degree of quality since most quality indicators show good-quality data.

Dataset 3: Tweets Dataset

Completeness–This dataset has a high degree of completeness as there is no missing data. Dataset has 766 records with 4 columns. Every column has 0% of null values, indicating that there are no missing values.

Consistency– This dataset is low consistent because there are special characters such as space, dash, and parentheses and Data values that might be interpreted in several ways.

Uniqueness–The dataset has duplicate values because conflicting duplications in different sources, for example the same data is in different ranks.

Integrity–There is no entity integrity because there is no primary key that uniquely identifies a record.

Conformity– All of the fields are in consistent form.

Accuracy– This dataset is low accuracy. There are misspellings in the dataset such as “feelin” instead of “feeling”.

Overall Quality – The dataset has a low degree of consistency, uniqueness, and accuracy. The overall quality of this dataset is not adequate, although there are some issues we might concern.

Dataset 4: Covid-19 Image Dataset

Completeness: The dataset contains 317 chest X-ray images, out of which 137 images are of COVID-19 patients and the rest are of patients with viral pneumonia or normal chest X-rays.

Consistency: The dataset seems to be consistent in terms of image format and labeling.

Uniqueness: The dataset is unique in that it contains chest X-ray images of COVID-19 patients, which may be valuable for research and diagnostic purposes.

Integrity: The dataset appears to be reliable and there are no known issues with the data integrity.

Conformity: The dataset conforms to standard image file formats (JPEG) and follows standard labeling conventions.

Accuracy: The accuracy of the dataset depends on the accuracy of the labeling and classification of the images.

Overall Quality: In general, there is hardly any missing data, the overall quality of this dataset is acceptable.

Dataset 5: Covid CXR Image Dataset

Completeness: The dataset includes a total of 1,823 chest X-ray images, which cover three different classes: Normal, Viral, and COVID-19.

Consistency: The dataset is consistent in terms of image size, format, and view, as all images are posteroanterior (PA) view of chest X-rays, in PNG format, and of size 299 x 299 pixels.

Uniqueness: The dataset does not provide information on whether there are any duplicate images, or how images were selected for inclusion in the dataset.

Integrity: The dataset does not provide detailed information on how the images were collected, labeled, or processed, which may raise questions about the integrity of the data.

Conformity: The dataset conforms to some common standards for medical imaging, such as the use of PA view chest X-rays and PNG format. However, there may be other aspects of conformity that are not addressed by the dataset, such as metadata or file naming conventions.

Accuracy: The accuracy of the dataset depends on how well the images represent the classes of interest (Normal, Viral, and COVID-19).

Overall Quality: In general, there is hardly any missing data, the overall quality of this dataset is acceptable.

2.5.1 Text Data Quality and Relevance Assessment

Quality Metrics

We examine not only the quality of the dataset, but also the quality of text. Assessing quality in a dataset context, such as missing values, duplicates, and expired data, and in a textual context, such as abbreviations, spelling mistakes, and grammatical errors. This task automatically assesses the data's quality. TextBlob is a Python library for handling textual data. If a sentence is misspelled, we can use the TextBlob.correct() method to find the correct words. We use three different functions in Pandas to detect missing values, duplicate values, and wrong data types. ISNULL () method takes a scalar or array-like object and indicates whether values are missing, returning a value if the expression is NULL. Duplicated () methods only analyzes duplicate values and returns a boolean series that is true for Unique elements dataframe.info() function is used to obtain a concise summary of the dataframe in order to determine data type. Regular Expressions (RegEx) is a syntax for matching a string to detect special characters and abbreviated words. Finally, language_tool_python.LanguageTool() is a Python library that detects grammatical and spelling errors in Python scripts. Table 1 contains a summary of data quality indicators. We restrict the proposed methods to textual data methods in the context of text analysis. We then categorize the metrics based on the quality dimension discussed in section 1.3 Research. Measurement criteria are determined by the number of indicators such as the accuracy indicator is Missing values, Misspelling, Ambiguous data, Incorrect and ungrammatical. To calculate the percentage of measured accuracy (Table2), sum the percentage of all indicators detected and divide by four to get a total accuracy’s percentage, the results are shown in Table3.

Table 2. Text data quality criteria

Data quality issue	Data Quality Dimension			
	Accuracy	Consistency	Uniqueness	Completeness
Missing data	x	-	-	x
Misspelling	x	-	-	-
Ambiguous data	x	x	-	-
Incorrect data and ungrammatical	x	x	-	x
Duplicated	-	-	x	-
Wrong data type	-	x	-	-
Special characters	-	x	-	-
Total%	25%	25%	25%	25%

Table 3. The percentage of each quality indicators

Dataset	Quality Indicators							
	Missing data (%)	Misspelling (%)	Ambiguous data (%)	Outdated data (%)	ungrammatical (%)	Duplicated (%)	Wrong data type (%)	Special character (%)

BBC News	0	0	7.36	0	0	5.96	20	7.34
Tweets	6.30	8.16	10.88	0	3.52	79.8	25	4.67
Mitre	34.54	0	0	0	1.75	72.11	6.25	2

Table 4. Text data quality assessment

Quality Dimensions	MITRE	BBC NEWS	TWITTER
Accuracy	90.93%	98.16%	92.79%
Completeness	81.86%	100.00%	95.09%
Uniqueness	27.89%	94.04%	20.20%
Consistency	97.50%	93.16%	88.98%

Relevance Metrics

We used the BERT Similarity Score to calculate the relevance metric. It measures the semantic similarity between sentences. Semantic similarity refers to the task that compares the similarity between one text to another. BERT is a massive corpus of trained data. It identifies the context of the sentence using Transformers. We applied a similarity score to extract covid keywords from documents using sentences. This metric emphasizes an important aspect of topic coherence measures which its value is between 0-1. It is determined not only by the topic, but also by the dataset used as a reference. The coherence metric is based on corpus probabilities, which are calculated as the number of documents containing the keyword divided by the total number of documents. The keywords used in the relevance calculation is Covid, Covid-19, 2019-nCoV, and Coronavirus.

Table 5. Relevance Scores

Dataset	BBC News	Tweets	Mitre
Relevance Scores	0.24	0.58	1

2.6 Other Data Sources

2.6.1 Tweets Blogs News – Swiftkey Dataset 4 million

This dataset consists of millions of tweets, blog posts, and news stories written in multiple languages. The texts have a variety of sentences and discussions that we frequently use in daily life. It contains 4 million texts totaling around 1 GB. Initially, we considered working on this dataset. Still, we found it challenging to frame business questions for this dataset as it has multiple languages: English, French, and Dutch. We found other datasets are the best fit for our project objective, so that's the reason we chose a different dataset.

2.7 Storage Medium

The storage medium for all the five chosen datasets can be any reliable and durable storage device that provides enough storage capacity to securely store the dataset. Due to the huge volumes of data being processed, a large external hard drive or a cloud storage service such as Amazon S3 or Google Cloud Storage may be suitable options. The flickr image dataset has over 100 GB of data. The Consumer Complaint Database has over 1.5 million records to explore. The BBC dataset has information coming in every day and it self-updates using a Python library and html. The COVID Synthea Dataset is quite large, with over 1 TB of data and the Denoising Dataset with Multiple ISO Levels has about 30 GB of data. So, in short, the storage medium selected for the datasets are quite the same.

2.8 Storage Security

Storage security is the implementation of controls, both physical and technical, to protect the stored data. For a storage security system to be viable, we will ensure four requirements. The first is confidentiality, the system must be able to protect against any unauthorized access from those wishing to steal or destroy the data held within it. The second is availability, the system must also be able to easily provide that information to authorized users. The third is backup and recovery, the system must have enough storage space to accommodate the backup procedure for data damaged by malware or ransomware. Finally, data integrity means ensuring that data cannot be altered or changed.

2.9 Storage Cost

All the chosen datasets occupy a combined space of around 200 GB. So based on the storage mediums discussed in section 2.7, we consider the cost associated with storing that amount of data should be quite affordable. We might have to invest in an external hardware drive or subscribe to a cloud service to ensure that the datasets are stored securely. The storage costs are to be covered under the AWS allotment for our project, so there are no costs incurred. However, S3 intelligent - tiering would cost about $8 per month for the 200 GB storage we would use. Since 3 teams are performing interrelated tasks, then we have an option to share the costs.

Section 3: Algorithms & Analysis / ML Model Exploration & Selection

3.1 Solution Approach

The goal of this project is to find out the ways to reduce the manual intensive steps in data management lifecycle functions, especially tagging data with their metadata. We will be using algorithms Topic Modeling as a part of Natural Language Processing; we used it for extracting covid related words and tagging covid related words. Topic Modeling automatically analyzes texts to determine cluster words. It does not require any labels or historical data that was previously classified by humans. A topic modeling algorithm can identify whether a set of incoming documents are receipts, complaints, advertisements, or such based on the content. Though it is like clustering, topic modeling algorithms identify the relationship between items. A dataset's hidden structure can be uncovered. We have used one specific statistical algorithm latent Dirichlet allocation or LDA from Topic modeling.

3.1.1 Systems Architecture

Before going into systems architecture, it is important to know the business goals. Out of the three text datasets, we are finding out which dataset gives better information about Covid-19. Similarly, we are finding which image dataset is better for classifying covid images. This model gives the user/business enough information to choose the most fit data for their business needs.

The schematic flow is depicted below in figure 10. The first step explains how we collect structured, semi-structured and unstructured data from publicly available sources and detect the data types as text and image. Both datasets are similar as they contain information and images about covid-19. They differ in the processing algorithm and the metrics used to describe their quality.

For text datasets, covid keywords are detected. Then the cleaned datasets are ingested into Momentum AI and predicted topics are generated using the pre-trained zero-shot classification algorithm. Then, data quality metrics are calculated, and the datasets are ranked in the order of highest relevance.

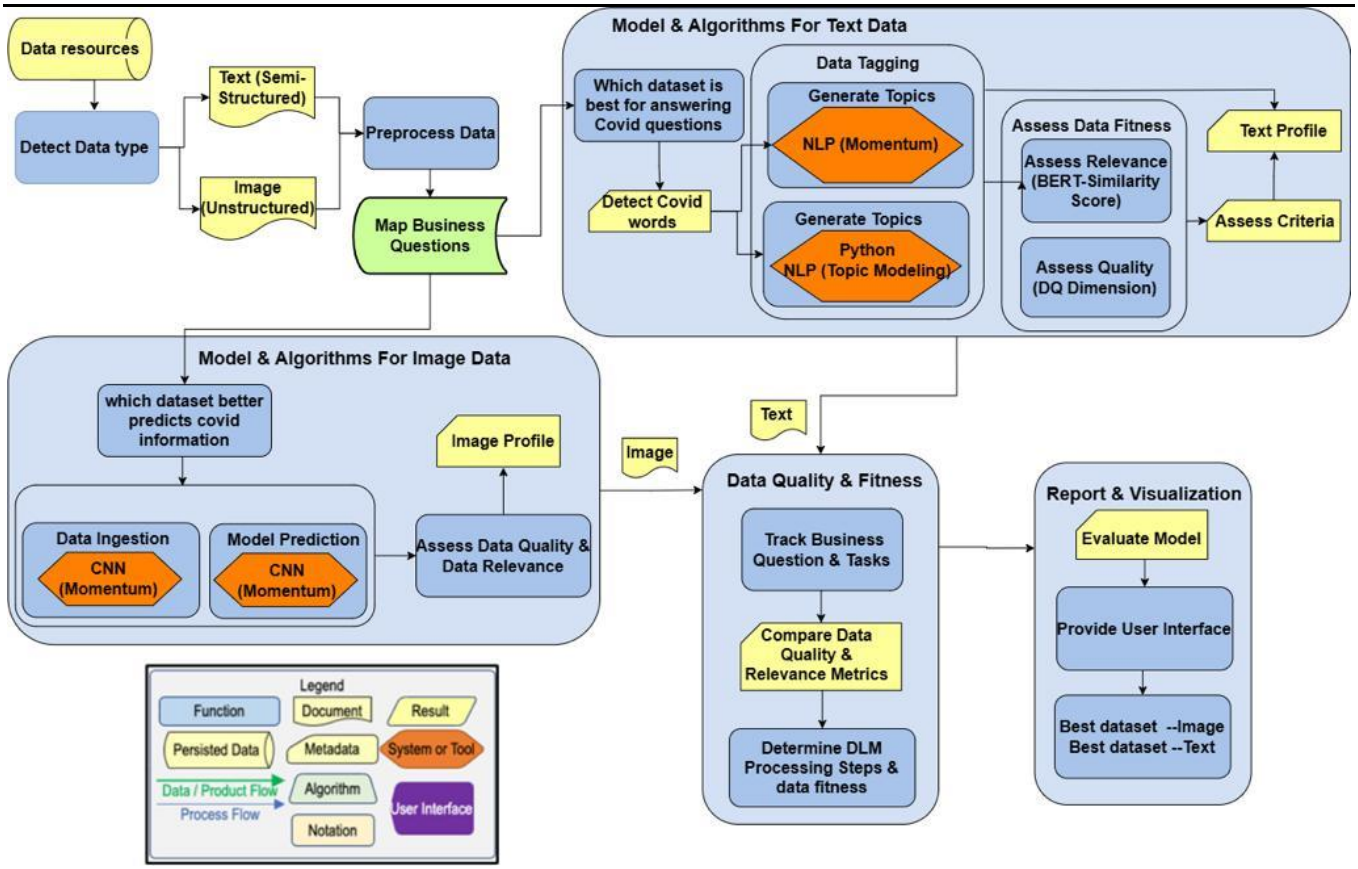

Figure 10: Schematic Flow

For image datasets, we generate metadata and profile images using CNN for image classification. Then we compare data quality and relevance metrics using Momentum AI. In the final step, after profiling images and text, we map business questions to generate fitness indicators and use momentum to map algorithms for each dataset. Finally, we combine the DLM process steps to generate reports and visualizations.

3.1.2 Systems Security

System security describes the controls and safeguards that an organization takes to ensure its networks and resources are safe from downtime, interference, or malicious intrusion. We use AWS Identity and Access Management (AWS IAM) to enable access control and permissions across multiple accounts.

Access control

Access control mechanisms are important to make sure that only authorized users can access the data, depending on how sensitive the image/text datasets are. This may involve taking steps like encryption or password security.

Backup and recovery

To ensure the image dataset is safeguarded in the event of a disaster or data loss, it's crucial to have a backup and recovery plan in place. This may involve routine data backups to a distant location and a strategy for recovering the data in an emergency.

3.1.3 Systems Data Flows

Input raw structured, unstructured, and semi-structured data from public data sources into a data type detection process that extracts image and text data. Text data inputs in the Gensim module include Latent Dirichlet Allocation (LDA) and Topic Modeling processes to detect related topics. The outputs include relevant terms extracted from text and their frequency in the dataset. Image data inputs in Convolutional Neural Networks (CNN) and Yolo processes to detect the accuracy of the dataset. The outputs include accuracy values of each image and the best-suited model for the business question.

3.1.4 Algorithms & Analysis

Text Datasets Algorithm

To validate data fitness, pandas and the natural language toolkit libraries are used on the text datasets in Jupyter Notebook. After the initial data preparation steps, the attributes containing sentences/phrases are extracted and combined into a single big paragraph using the sum function. Then the individual words are converted to lower cases and tokenized using the Regular Expression module. These sets of words also contain multiple instances of stopwords such as “a”, “the”, “of”, “as”, etc. that do not add meaning for the analysis. These are removed before continuing towards the next steps. Now covid related words such as booster, covid, vaccine, etc. are searched from the whole corpus using another customized function. Once those are collected, the number of times those instances occur are found using the frequency distribution function. For example, in the BBC News dataset, “covid” appears 151 times and “booster” appears 11 times. Then using the matplotlib library, the top 20 frequently occurring words are visualized. Similarly, the word frequency counts are calculated for MITRE Synthetic Health Data and Tweets datasets.

To perform data tagging, the cleaned datasets are modeled in Momentum AI - a no code/low code platform for machine learning and data engineering tasks. Each dataset is first uploaded, ingested, then trained using the pre-trained zero-shot algorithm. The trained model is then used to make predictions based on the ingested datasets. As a result, labels and scores are generated for each predicted topic. Then we proceed to use the built-in SQL interface to query the distribution of each topic.

Metadata is generated for comparing quality metrics using KPI to set up thresholds for each quality dimension. The other way to compare relevance metrics is through detecting text and clean data by removing punctuation and stop words and then comparing relevance metrics using Topic modeling and Named Entity Recognition (NER) in NLP. Topic Modelling and Named Entity Recognition is used to extract topics from large volumes of text for tracking the business questions. Here 'topic' is the business questions keywords. Latent Dirichlet Allocation (LDA), which is part of Python's Gensim package, is used for Topic Modeling. The two main inputs to the LDA topic model are the dictionary and the corpus. A dictionary for the business questions is created. The dictionary object from Gensim is applied, which maps each word to their unique keywords for business questions.

Image Datasets Algorithm

The first stage is to perform the preprocessing of the data. The images are of different dimensions. With Momentum AI, the images are resized to 240X240 dimensions for fast computation using the built-in Convolutional Neural Networks. Then the model is run through CNN layers. The dataset which has more accuracy and better data quality metrics is the best fit model for the business question. After this stage the metadata is extracted for the images and the tags are generated for each image.

3.2 Machine Learning

3.2.1 Model Exploration

Latent Dirichlet Algorithm is an unsupervised machine learning algorithm, which is a part of the Gensim package. There are 5 steps involved in this modeling process. In step 1, data is imported into a pandas data frame. Next is the preprocessing stage, where spaCy and NLTK are used. SpaCy's pre-trained model acts as a pipeline for tokenizing the text. Parts of Speech (POS) tags are assigned to each word to differentiate each token as an adjective/verb/punctuation, etc. Stop words are removed by providing the unwanted POS as a list. Once this is done, a clean list of tokens is obtained. In step 3, the dictionary and corpus are created, which are the major inputs to the LDA model. The "Dictionary" object from Gensim is used which maps each word to their unique ID. The words that occur with high and low frequencies are removed and then the corpus is created. Model building starts now. To train the unsupervised machine learning model on the data, LDA-multicore is used for faster parallel processing. The corpus is iterated about 50 times to optimize model parameters, number of topics is set as 10, and the workers to be 4. Once the model is trained, it is evaluated using a coherence score which ranges between 0 and 1. A score above 0.5 is good. The final step is to print and visualize the topics. The summary of the first sentence is compared to the LDA model to see the percentage of occurrence of each topic. Using pyLDAvis, the clusters are visualized. Once this is done, a new column is created in the data frame and the most probable topic is added to it.

CNN (Convolutional Neural Network) is a deep learning algorithm used for image detection and classification tasks. The algorithm is built on the principle that the input image is processed via several convolutional layers to extract significant features, then layers are pooled to minimize the spatial dimensions, and eventually the output is passed through fully connected layers for classification. The first step involved in CNN is the data preparation dataset must first be set up for training. Images must be gathered, labeled, and the data divided into training and validation sets. The fundamental component of a CNN is the convolutional layer. To extract features from the input image, it applies a collection of filters (sometimes referred to as kernels). Each filter's output is a feature map, and the convolution layer's output is created by stacking these feature maps. To add non-linearity to the model, an activation function is used to the convolution layer's output. ReLU (Rectified Linear Unit) and Sigmoid have been used. To minimize the spatial dimensions of the feature maps generated by the convolutional layers, pooling layers are used. As a result, the model's parameter count is decreased, and overfitting is avoided. The output of the convolutional and pooling layers is sent through fully connected layers.

To create a probability distribution over the classes, these layers apply a set of weights to the input and generate a set of output values. These output values are then processed via a SoftMax function. The loss function computes the difference between the true label of the input image and the expected output of the model. A separate test set is used for evaluation. For image categorization, the metrics are accuracy, precision, recall, and F1-score. Once the model is trained and assessed, it can be used to make predictions about new, unviewed photos. The model processes the input image, and the output of the SoftMax function is used to guess the image's class label.

3.2.2 Model Selection

The machine learning models were selected after carefully analyzing their pros and cons. For text datasets, LDA model is chosen for topic modeling. It is how hidden topics of the same theme are extracted from a document. The frequency of the words in the topic and their distribution are the outputs. LDA performs faster classification, easier to implement and quick in terms of runtime, allowing it to handle large datasets. It is also a visualization and interpretation tool. But LDA's process treats documents as a set of words only and ignores syntactic and semantic information. It is necessary to fix the number of topics. Dirichlet topic distribution cannot capture correlation.

For Image datasets, the algorithm used is based on Convolution Neural Networks (CNN). It is highly accurate in image detection, frequently outperforming other machine learning methods. It also supports automatic feature learning. The necessity for manual feature engineering is eliminated because of CNNs' ability to automatically learn and extract pertinent features from input images. Image detection tasks can be started with pre-trained CNN models, making training quicker and more effective. But the downside is, to learn and generalize effectively, CNNs need a lot of labeled training data. This can be time-consuming and expensive to obtain. CNNs require a lot of computing power and must be trained and tested using GPUs and other sophisticated hardware. CNNs are susceptible to adversarial attacks, wherein tiny alterations to an input image can result in incorrect categorization.

Section 4: Visualizations / ML Model Training, Evaluation, & Validation

4.1 Overview

In the previous sprint, the solution approach and algorithm selection were finalized for the image and text datasets. While Convolution Neural Networks is used for image analysis, topic modeling is done for analyzing texts and extracting data quality metrics. In this sprint, the results are observed and visualized, and the machine learning models are evaluated and checked if justifiable.

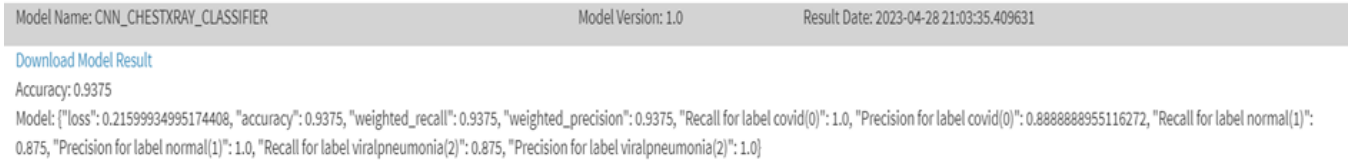

Figure 11: Results for Covid-19(Image Dataset)

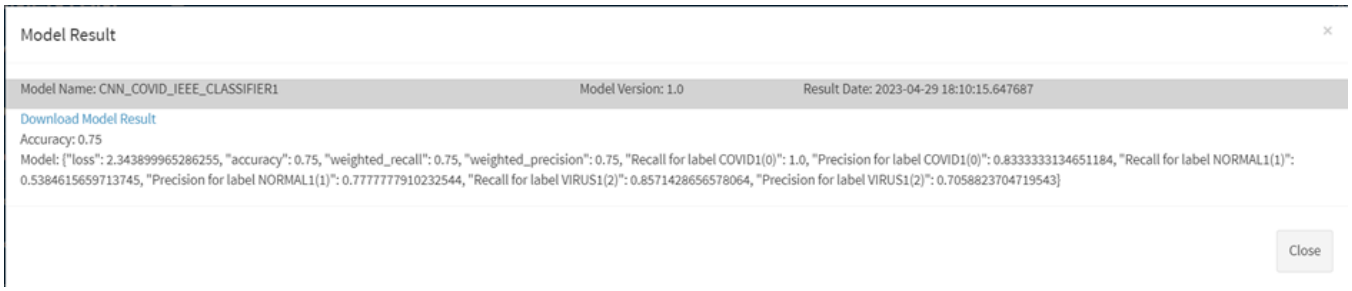

Figure 12: Results for Covid CXR (Image Dataset)

4.2 Visualizations

4.2.1 COVID-related words Word Frequency

One of the first steps in natural language processing (NLP) is the ability to count the frequency of words used in text documents. Because our goal is to determine which dataset is best for COVID, we examined the number of COVID-related words in the dataset to determine the dataset's relevance to Keywords.

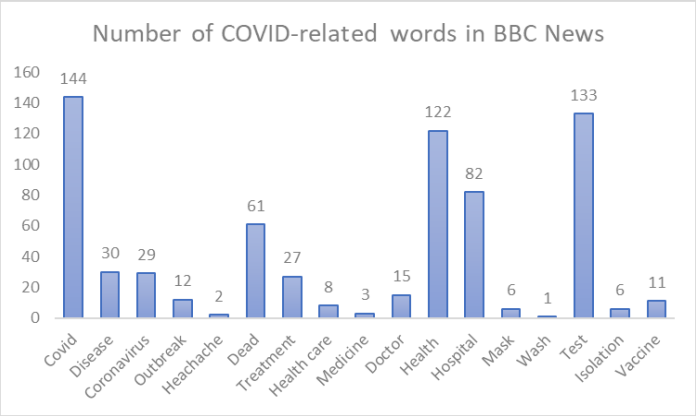

Figure 13: Covid World in BBC News

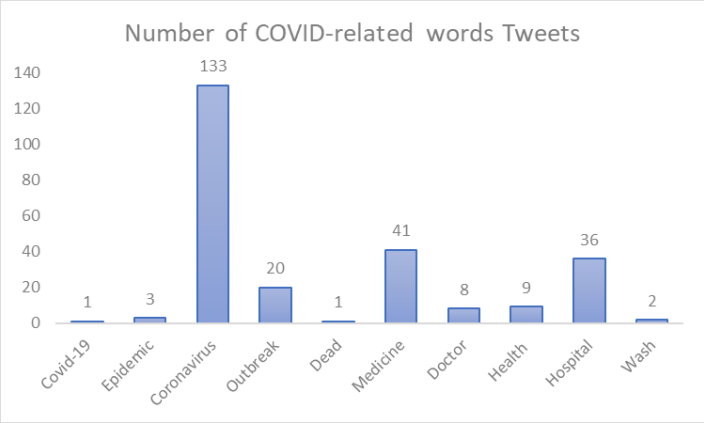

Figure 14: Covid World in Tweets

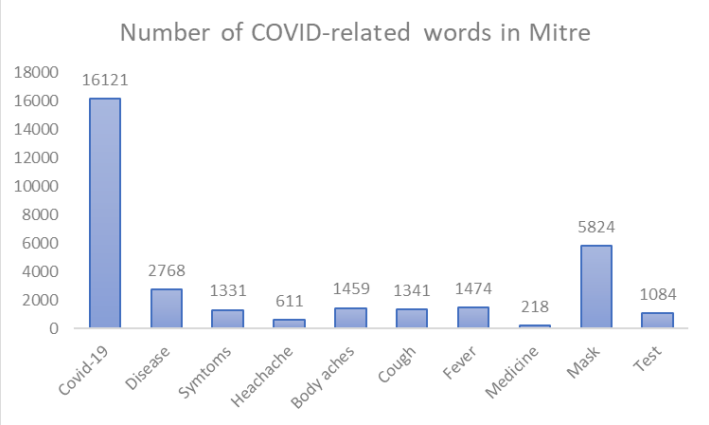

Figure 15: Covid World in MITRE COVID19 100K

4.2.2 Words Frequency Counts

Word frequency count is a quick way to explore the distribution of frequent words in a corpus. The below images show the top 20 commonly occurring words in Tweets and BBC datasets.

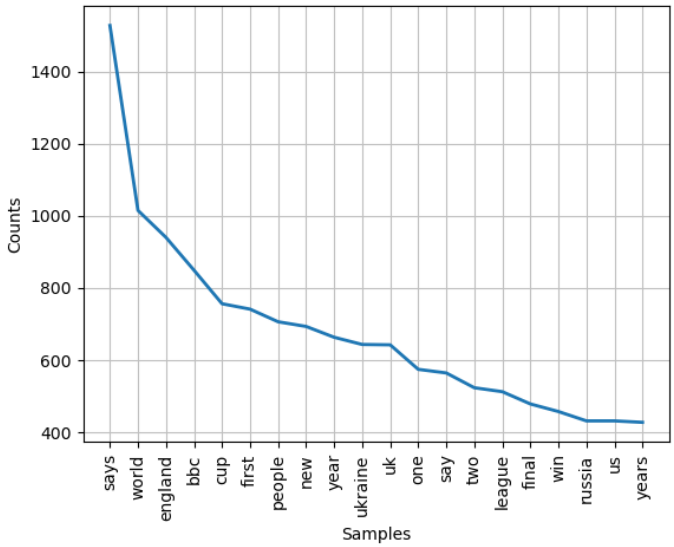

Figure 16: Frequent words in BBC

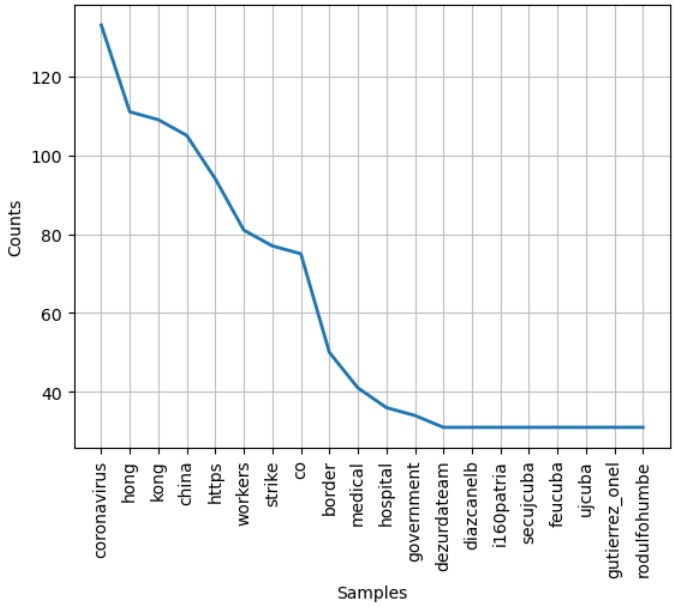

Figure 17: Frequent words in Tweets

4.2.3 Word Clouds

To provide a visual representation of the common texts present in the datasets, the word cloud is created. These word clouds give us an estimate about what the text in the dataset looks like and what it talks about. In this visualization, the higher the frequency of the word, the larger it appears. Even though the word cloud does not show the tags, any unstructured text data can be quickly visualized to reveal patterns and trends.

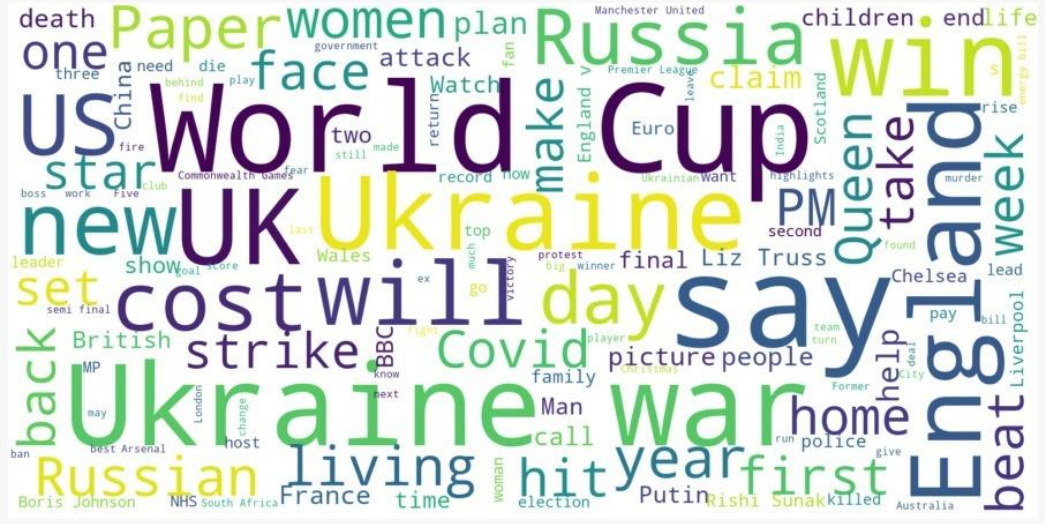

Figure 18: World Cloud of BBC News

4.2.4 Topic Modeling for BBC News Dataset

The optimal number of topics generated by the LDA model for all datasets is determined to be 10. These topics are distinct, segregated, and organize the dataset in the best way possible. These visualizations are excellent for summarizing datasets by displaying words, topics, and contents.

After putting through the BBC datasets in the LDA model, the optimal number of topics is found to be 10. The following figures show the intertopic distance map.

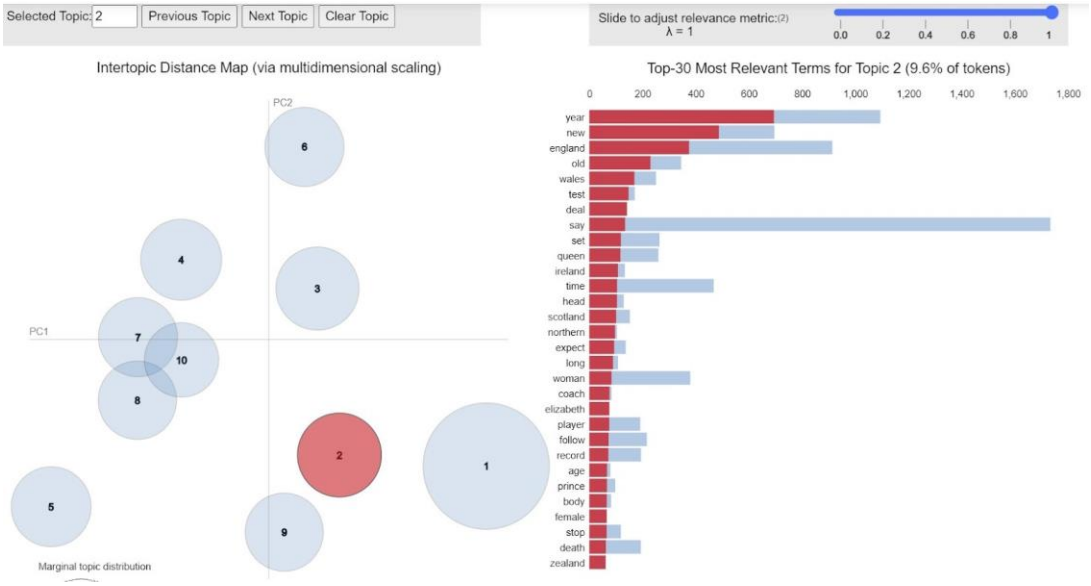

Figure 19: Intertopic Distance Map - BBC

The subplot visualization below as shown in figure 21 shows those topics (from 0 to 5) along with their keywords. The keywords are represented as overlapping bar graphs with word counts and weights. It counts the number of times a keyword appears in text and assigns weight to each keyword based on its importance.

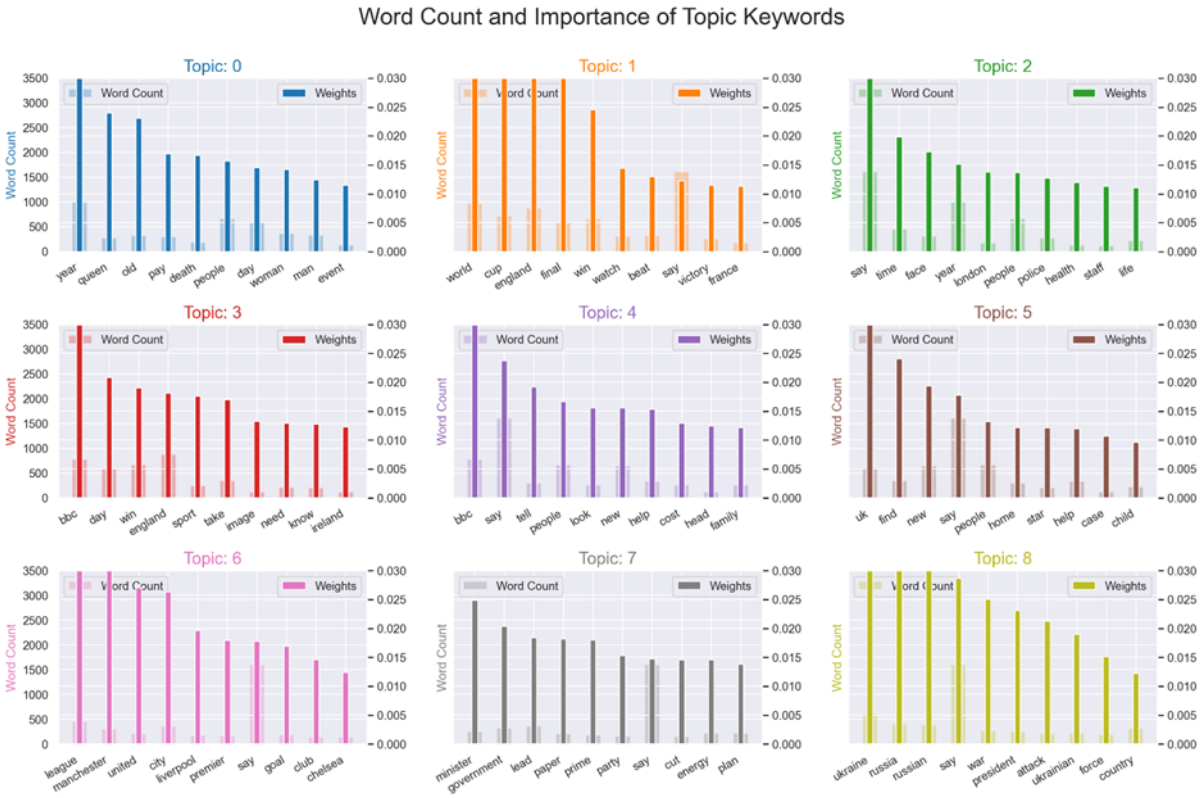

Figure 20: Word Count and importance of Topic Keywords

The visualization below shows that the BBC dataset consists of a variety of topics (Figure 21). The keyword Topic 0 appears to point to "Identification". Topic 1 has a keyword related to "Competition", while the others include "Health", "Sports and Entertainment", "People", and "England".

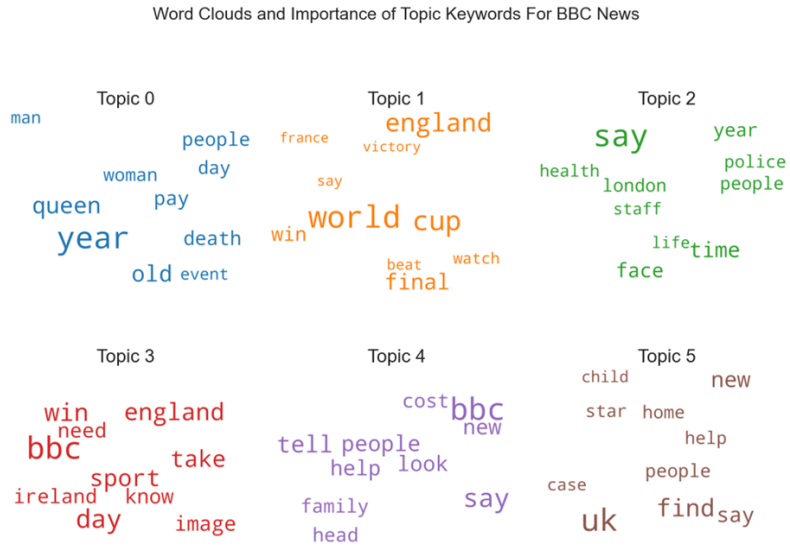

Figure 21: Word Clouds of Topics based on Keywords.

4.2.5 Topic Modeling for Tweets Dataset

After putting through the Tweets datasets in the LDA model, the optimal number of topics is found to be 10. The following figures show the intertopic distance map.

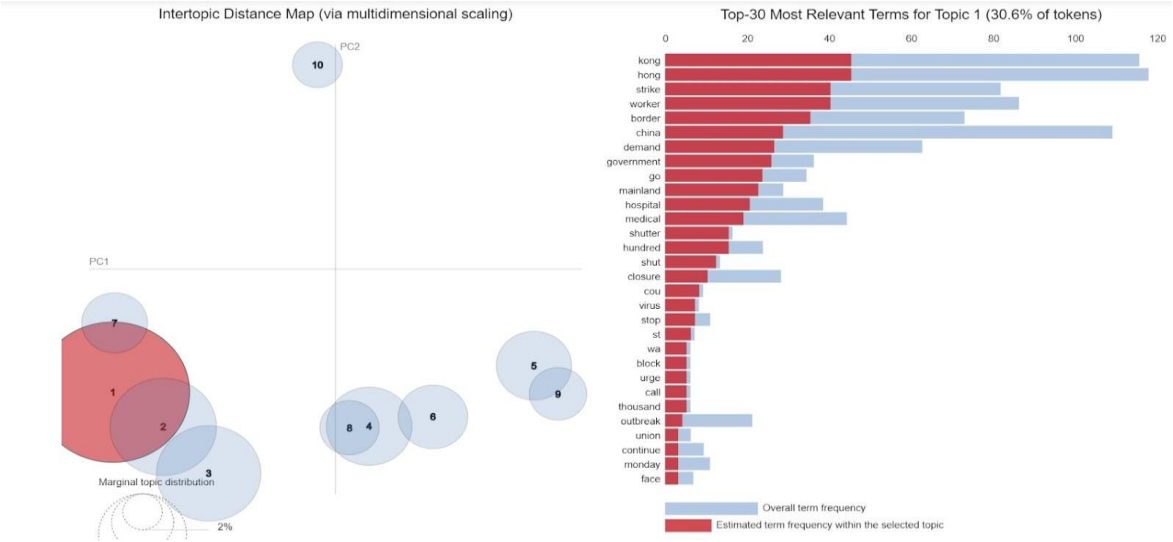

Figure 22: Intertopic Distance Map - Tweets

These topics, along with top keywords, are shown in the image below (Figure 23). The top six topics, as well as the top ten keywords in each topic. Furthermore, all concordance scores were extracted from the corpus fed into the LDA.

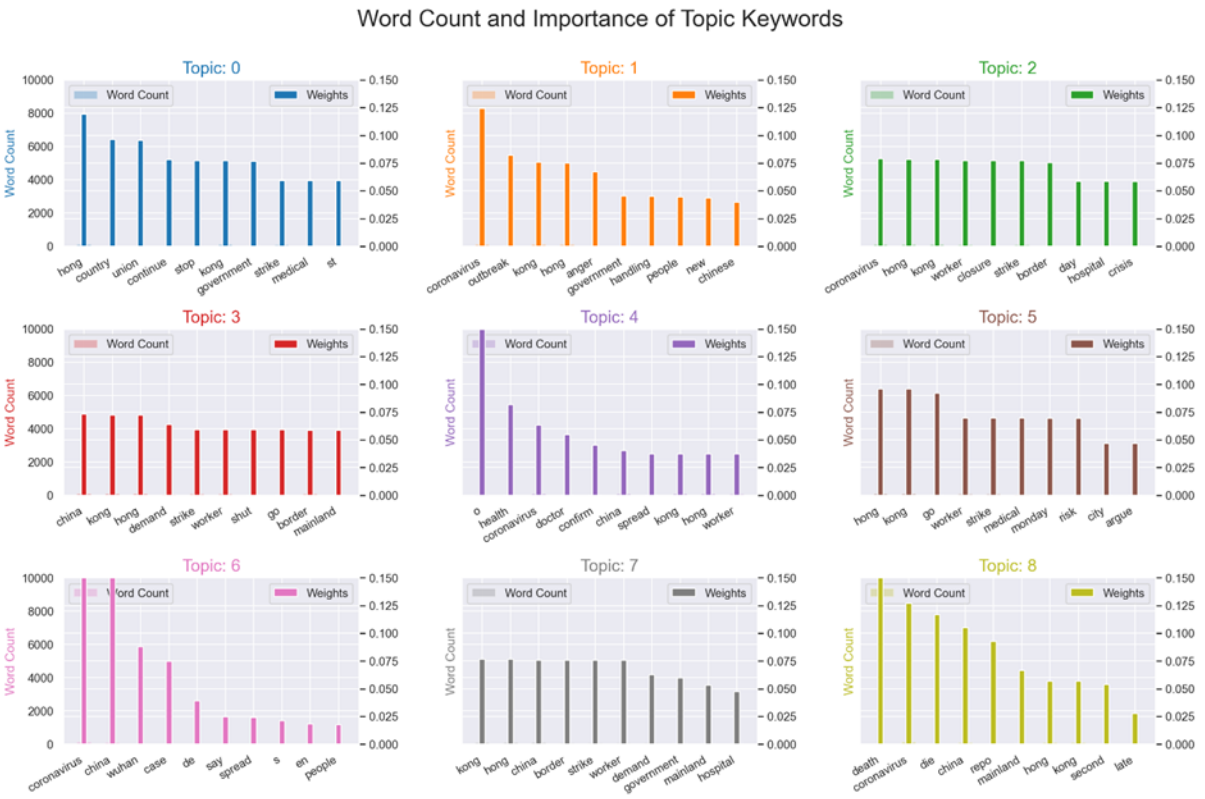

Figure 23: Word Count and importance of Topic Keywords

The main topics found in the Tweets dataset are 'government', 'coronavirus', 'politics', 'hospital', 'demand', 'health' and 'workers'. From this, it can be inferred that most of the tweet dataset is talking about the coronavirus pandemic and its impact. Keywords are displayed in different colors and the larger the keyword, the more weight it has.

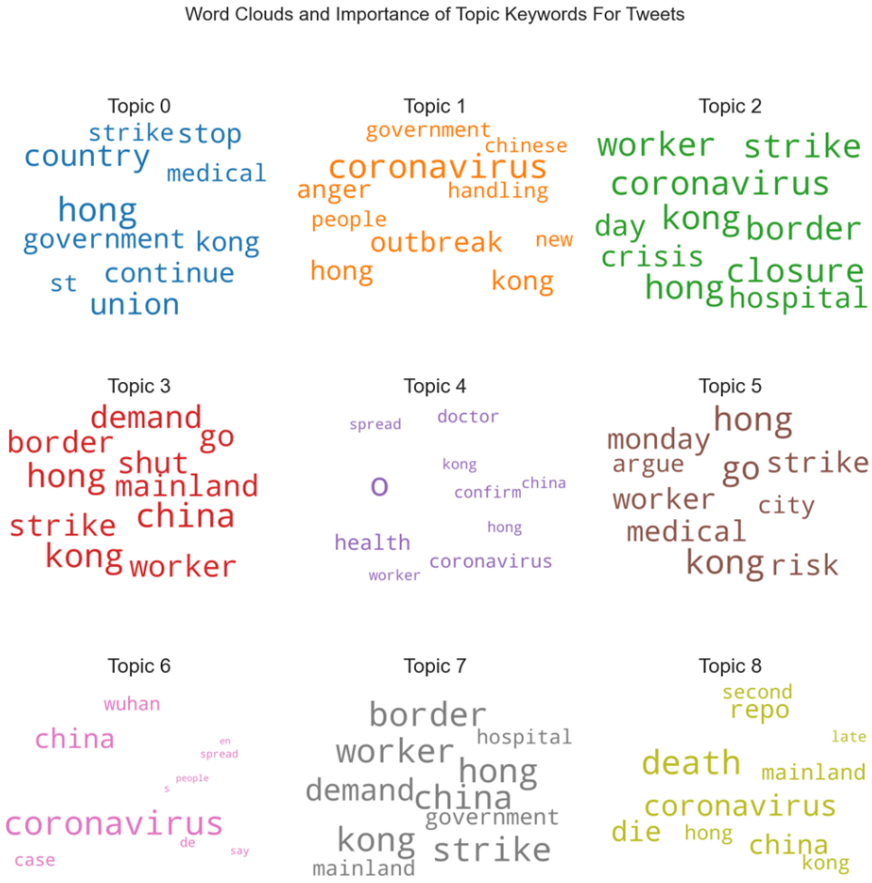

Figure 24: Word Clouds of Topics based on Keywords

4.2.6 Topic Modeling for MITRE COVID19 100K

After putting through the MITRE dataset in the LDA model, the optimal number of topics is found to be 10. The LDA corpus input was used to extract the top six topics and top ten keywords within each topic. The keywords that were most like each other were grouped together under one topic. Each keyword is represented by two parameters: its relative frequency in this dataset and its weightage as shown in figure25.

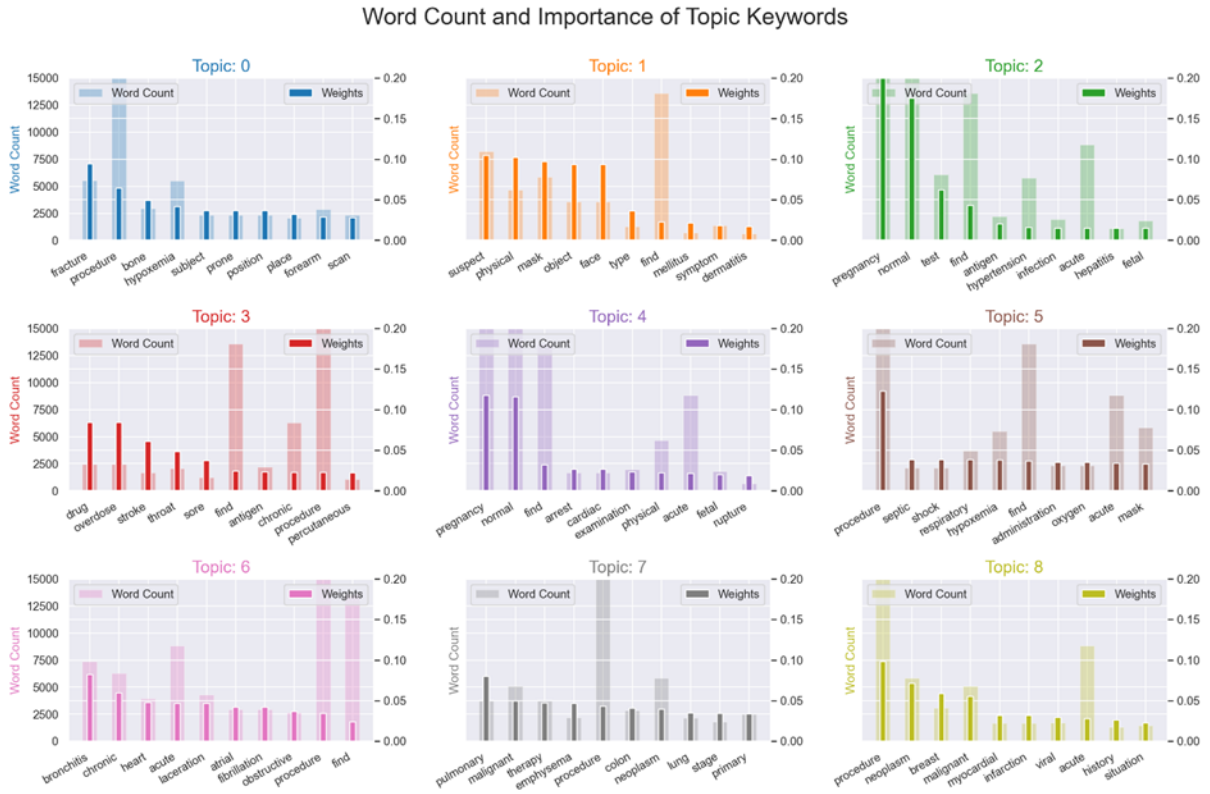

Figure 25: Word Count and importance of Topic Keywords

The main topics found in the Mitre dataset are 'procedure', 'coronavirus', 'physical', 'condition', 'symptom', ‘medicine' and ‘treatment’. From this, it can be inferred that most of the Mitre dataset is talking about medical and patient.

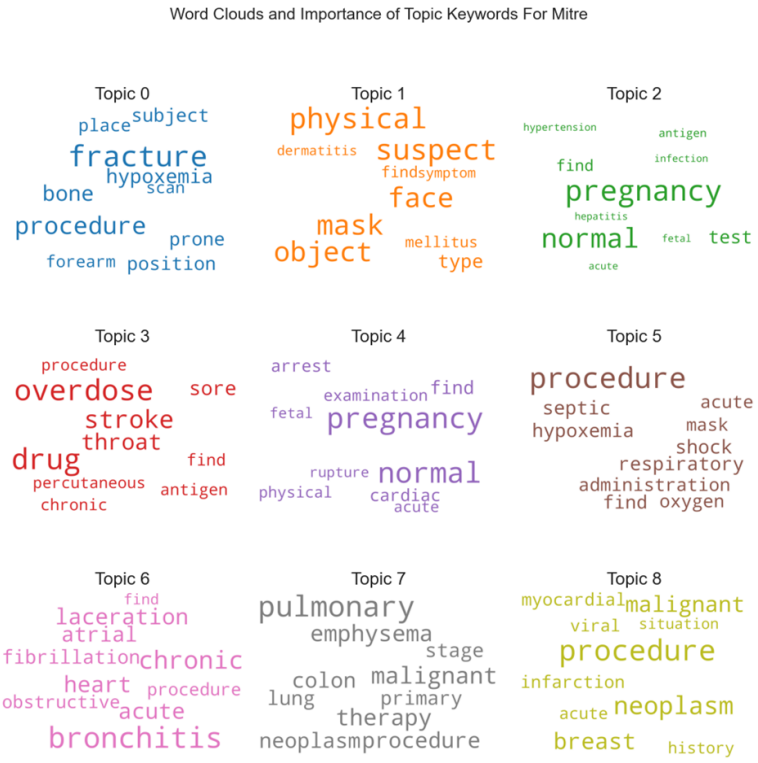

Figure 26: Word Clouds of Topics based on Keywords.

4.2.7 Visualizations for Image Datasets

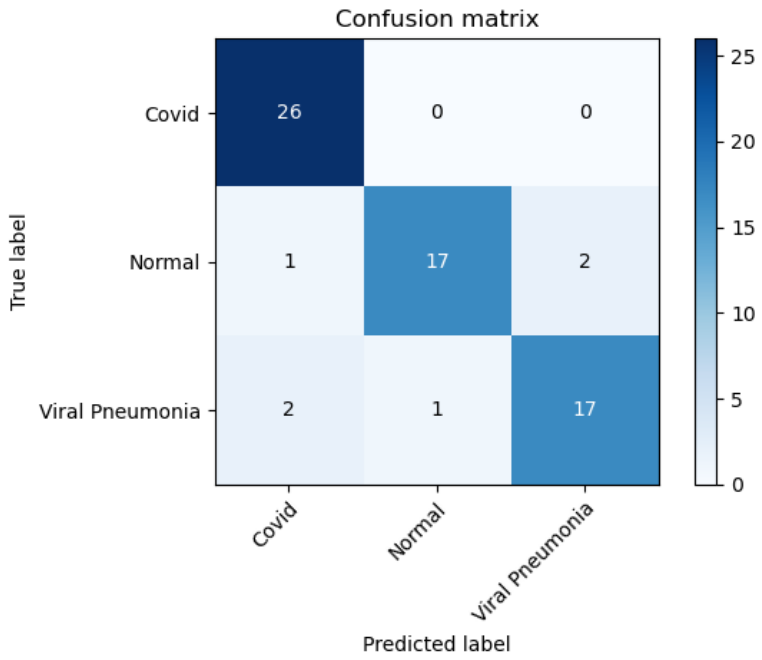

Figure 27: Confusion Matrix Covid-19(Image Dataset)

The above graph represents the results of a classification model that has been trained to distinguish between three different classes: Covid, Normal, and Virus. The X-axis of the graph represents the actual labels of the data points, while the values represent the predicted labels by the model. The values in the graph represent the number of data points that fall into each category. From the graph, we can see that the model correctly predicted most of the data points. Specifically, it correctly classified 26 out of 26 Covid cases, 17 out of 20 Normal cases, and 17 out of 20 Viral Pneumonia cases. However, there were some misclassifications, for instance, the model predicted 1 Normal case as Covid and 2 Normal cases as Viral Pneumonia. It also predicted 2 Viral Pneumonia cases as Covid and 1 Viral Pneumonia case as Normal.

Figure 28: Loss vs Iteration for Covid-19(Image Dataset)

The above graph is about Loss versus Iteration for Covid-19 Image dataset. As the iterations increase the loss decreases, which is better for the model performance.

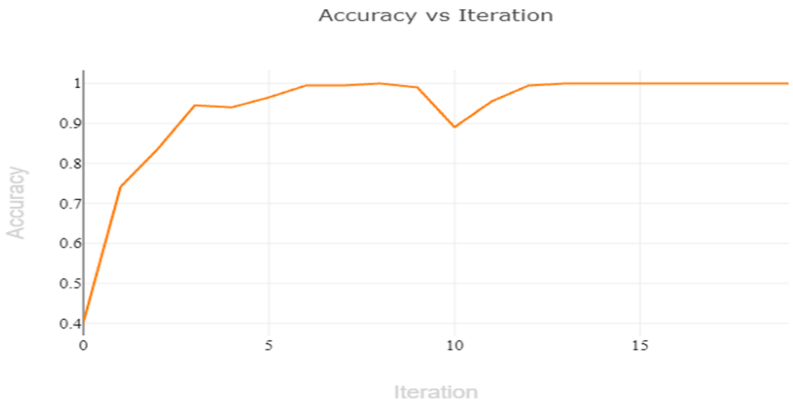

Figure 29: Iteration vs Accuracy for Covid-19(Image Dataset)

The above graph explains about iteration against accuracy plot. As we can observe that more the iterations the better the accuracy for the model is which ultimately is a good indicator that model is good.

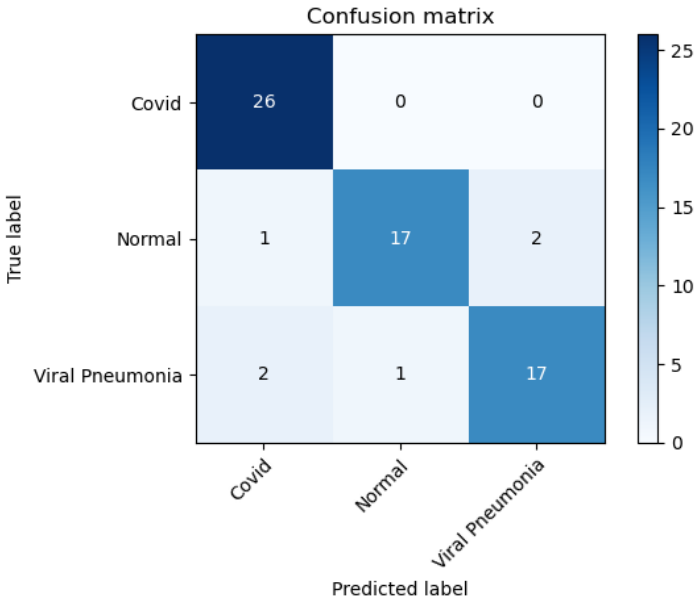

Figure 30: Confusion Matrix Covid CXR(Image Dataset)

The above graph represents the results of a classification model that has been trained to distinguish between three different classes: Covid, Normal, and Virus. The X-axis of the graph represents the actual labels of the data points, while the values represent the predicted labels by the model. The values in the graph represent the number of data points that fall into each category. From the graph, we can see that the model correctly predicted most of the data points. Specifically, it correctly classified 347 out of 347 Covid cases, 624 out of 667 normal cases, and 605 out of 616 virus cases. However, there were some misclassifications, for instance, the

model predicted 16 normal cases as Covid and 3 virus cases as Covid. It also predicted 8 virus cases as normal and 27 normal cases as virus.

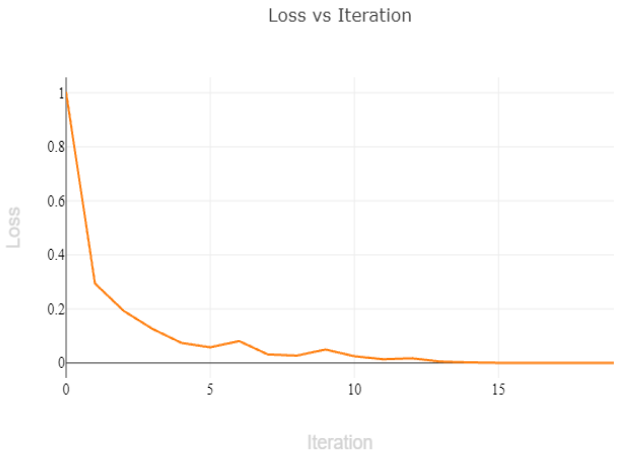

Figure 31: Loss vs Iteration for Covid CXR (Image Dataset)

The above graph is about Loss versus Iteration for Covid CXR Image dataset. As the iterations increase the loss decreases, which is better for the model performance.

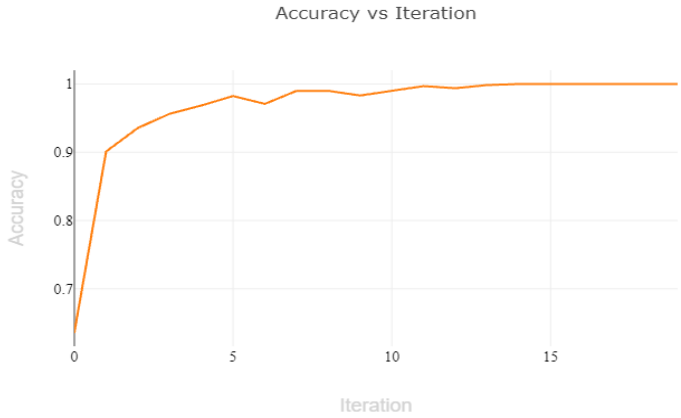

Figure 32: Iteration vs Accuracy for Covid CXR (Image Dataset)

The iteration against accuracy plot graph shown is illustrated above. We can see that the model's accuracy improves as the number of iterations increases, which is a good sign that the model is sound.

4.3 Machine Learning

Latent Dirichlet Algorithm (LDA) is the topic modeling method used here. LDA classifies text in a document to a particular topic. It has two parts - the words that belong to a document which we already know, and the probability of words belonging to a topic which needs to be calculated. The number of topics is chosen beforehand.

Convolutional Neural Network (CNN) is a renowned deep learning technique used for image identification and classification applications. The approach is predicated on the notion of processing the input

image via numerous convolutional layers to extract relevant information, followed by layer pooling to minimize the spatial dimensions, and finally sending the output through fully connected layers for classification.

4.3.1 Model Training

The LDA multicore model from Gensim is implemented here. This parallelizes and speeds up model training. The training algorithm is streamed and runs in constant memory with respect to the number of documents. The corpus and dictionary are fed into the model along with the number of iterations and topics. Then the topics are printed, and the coherence score is calculated. This measures the degree of semantic similarity between high scoring words in the topic.

The CNN model is trained using the Adam optimization algorithm, which modifies the network's weights to reduce the loss function. The loss function calculates the difference between the model's predicted output and the actual label of the input image.

4.3.2 Model Evaluation

The performance LDA model is best described by two methods. The coherence score and perplexity provide a convenient way to assess the quality of a given topic model.

Perplexity measures how well the topic model predicts new or previously unseen data. It reflects the model's ability to generalize. A low perplexity score indicates that the model is confident in its predictions. A high perplexity score indicates that the model's predictions are uncertain and inaccurate. Coherence measures how well words in a topic relate to one another. It reflects human intuition about what constitutes a good topic. A high coherence score indicates that the subject matter is consistent, clear, and relevant. A low coherence score indicates that the subject is ambiguous or irrelevant. Coherence and perplexity which turn out to be –4.89 and 0.38 for this initial model, respectively as shown in Figure 33.

```
# Compute Perplexity Score
print('nPerplexity Score: ', lda_model.log_perplexity(corpus))

# Compute Coherence Score
coherence_model_lda = CoherenceModel(model=lda_model, texts=tweets['Tweet'], coherence='c_v')
coherence_lda = coherence_model_lda.get_coherence()
print('nCoherence Score: ', coherence_lda)
```

```
nPerplexity Score:  -4.890494378382867
nCoherence Score:  0.38229601086261633
```

Figure 33: Coherence and perplexity of the gensim model

4.3.3 Model Validation

Coherence scores for each number of topics for tweets dataset is shown in figure 35. We discovered that the highest coherence score was used to run the model at 3 different topics. To improve the LDA model's efficiency, we optimize it by determining the optimal number of topics. Number of Topics (K), Dirichlet hyperparameter alpha (document topic density), and Dirichlet hyperparameter beta (word topic density) are all hyperparameter adjustments. We run these tests sequentially, one parameter at a time while holding the others constantly, and over the two different validation corpus sets. We identify the function and iterate over the topics, alpha, and beta parameter values. In this case, we chose K=3, alpha = 0.01 and beta = 0.1, and this results in a 20.2% improvement over the initial score shown in figure 37. Table3, we summarized the efficiency gains of model tuning, which can increase model quality by up to 27.1%.

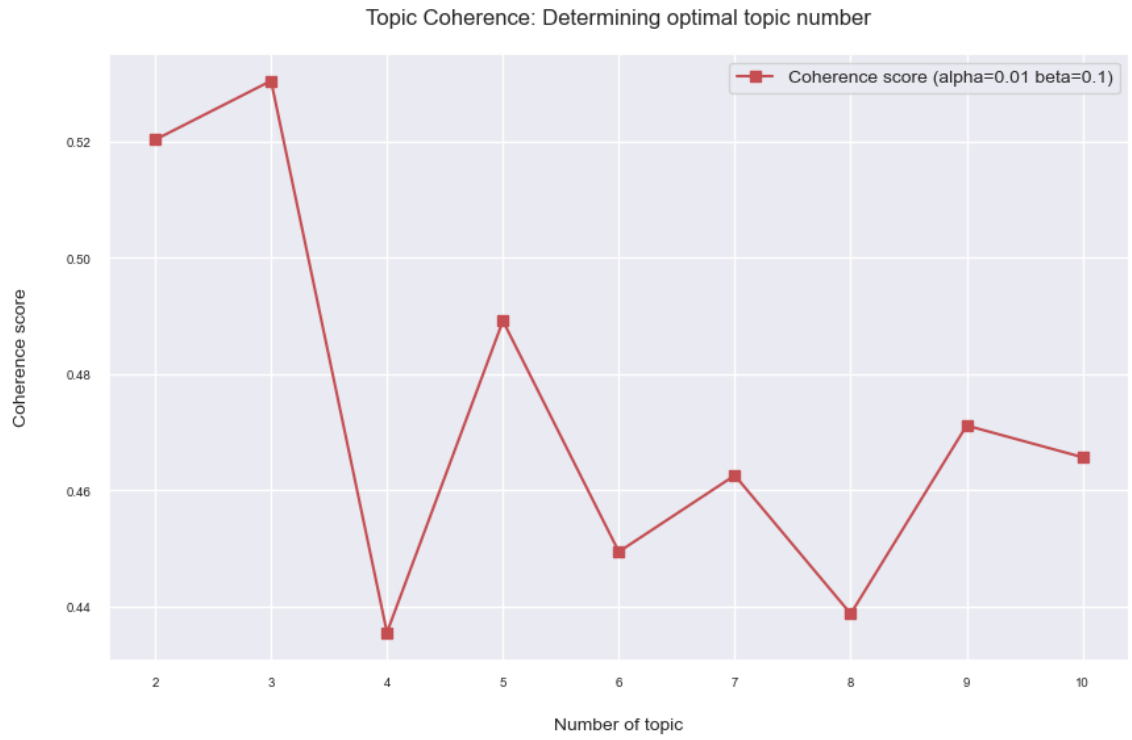

Figure 34: Coherence scores for each number of topics for tweets dataset

```
# Prepare the data for drawing chart
target_alpha = 0.01
target_beta = 0.01
topic_nums = list(range(2, 11))
target_co_pos_set = set()
for i, t in enumerate(zip(model_results['Alpha'], model_results['Beta'])):
    if t[0] == target_alpha and t[1] == target_beta:
        target_co_pos_set.add(i)

coherences = []
for i, co in enumerate(model_results['Coherence']):
    if i in target_co_pos_set:
        coherences.append(co)

for topic_num, coherence in zip(topic_nums, coherences):
    print("Topic number={} with coherence value={:.02f}".format(topic_num, coherence))
```

```
Topic number=2 with coherence value=0.52
Topic number=3 with coherence value=0.53
Topic number=4 with coherence value=0.44
Topic number=5 with coherence value=0.49
Topic number=6 with coherence value=0.45
Topic number=7 with coherence value=0.46
Topic number=8 with coherence value=0.44
Topic number=9 with coherence value=0.47
Topic number=10 with coherence value=0.47
```

Figure 35: Finding optimal number of topics for tweets dataset.

```
best_co = target_collection[0][2]
improve_pert = (best_co - coherence_lda) * 100 / coherence_lda
print("Coherence score is improved by {:.01f}%".format(improve_pert))
```

```
Coherence score is improved by 20.2%
```

Figure 36: Score increased after model tuning for tweets dataset.

Table 6. Performance Comparison before and after Tuning Topic Model

Dataset	*Initial Number of Topics*	*Coherence score(C_v)*	*Optimal Number of Topics*	*New Coherence score(C_v)*	*%Coherence score improvement*
BBC News	*10*	*0.33*	*5*	*0.42*	*27.1%*
Tweets	*10*	*0.38*	*3*	*0.46*	*20.2%*
MITRE	*10*	*0.32*	*8*	*0.40*	*25.0%*

Section 5: Findings

5.1.1 Topic Modeling - Results

Once topics were generated using zero-shot classifier, we used SQL commands in the Momentum interface to count the frequency of each prediction and arrived at getting the below metrics. Comparing figures 37, 38 and 39, we can infer that the MITRE dataset has the greatest number of covid-related words, since the number of topics predicted is higher.

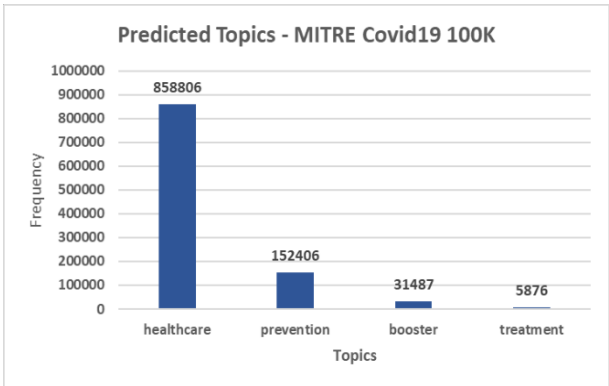

Figure 37: Topics Frequency - MITRE Covid19 100K

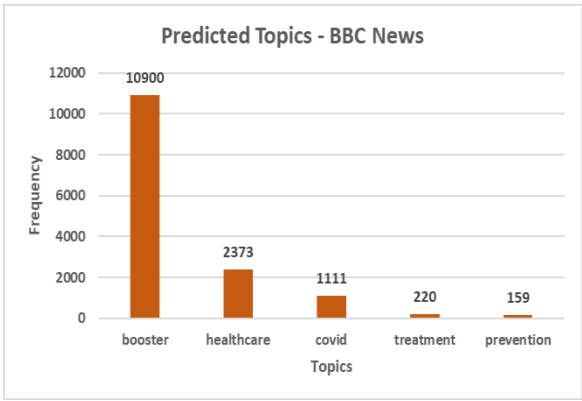

Figure 38: Topics Frequency - BBC News

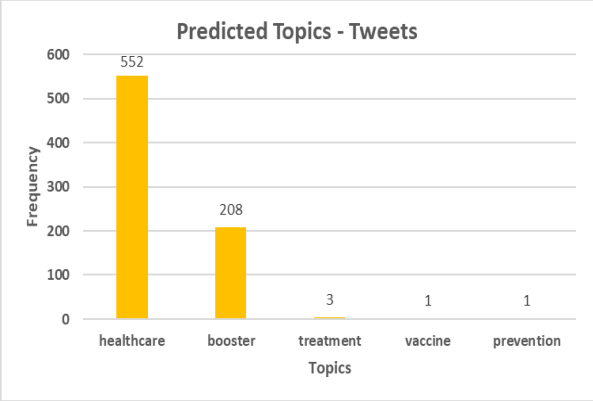

Figure 39: Topics Frequency – Tweets

5.1.2 Image Classification – Results

For both datasets, the accuracy is high, with Covid-19 Image dataset having an accuracy of 93.75%, and COVID CXR Image dataset having an accuracy of 84.37%. This indicates that the data in both datasets is reliable. Also, the F1 Score for Covid-19 Image dataset is 93.75%, and that of COVID CXR Image dataset is 84.37%. This indicates that the datasets are relevant for their intended purposes, which is the detection and diagnosis of respiratory diseases, including COVID-19

5.1.3 Relevance Metrics

Relevance metrics were executed in Jupyter Notebook and subtopics were generated. In Fig 37, we see that coherence score and similarity score are higher in MITRE when compared to the other two datasets. This means that the MITRE dataset provides more relevant information regarding covid topics or business questions.

Topic	Subtopic	Coherence score			Relevance score (Similarity score)		
		BBC News	Tweeter	MITRE	BBC News	Tweeter	MITRE
Topic1: Covid	['covid'', 'covid-19', '2019-nCoV', 'coronavirus']	0.24	0.58	1.00	0.22	0.23	0.40
Topic2: Symptoms	['symptom', 'headache', 'body aches', 'cough', 'dead', 'fear']	0.15	0.28	0.36	0.12	0.29	0.76
Topic3: Prevention	['social distance', 'mask', 'sanitizer', 'wash', 'test', 'isolation']	0.26	0.23	0.08	0.25	0.32	0.46
Topic4: Treatment	['healthcare', 'medicine', 'doctor', 'health', 'hospital', 'treatment']	0.30	0.34	0.34	0.17	0.32	0.68

Figure 40: Comparison of Relevance Metrics

Table 7: Comparison of metrics - Text

Dataset	Business question	Data quality score	Relevance score	Fitness score
MITRE	Most relevant	86.34%	80%	83.17%
BBC NEWS	Least relevant	95.65%	44%	69.83%
TWEETS	More relevant	85.76%	46%	64.88%

We found that the MITRE dataset to be the most fit for our business question with 84.55% quality and 80% relevance, so the overall fitness score is 82.28%, followed by BBC and Tweets with 69.83% and 62.64% respectively. Furthermore, after we got the results of Topic modeling, we found that the BBC news dataset contained mostly sports and competitive topics. Tweets dataset is mainly about coronavirus pandemic and its impact. Lastly, MITRE dataset covers most medical and covid patients.

Table 7 above gives a comprehensive score of data quality, relevance, and fitness. So even though BBC News has the highest percentage in terms of quality, its fitness to our business question is only secondary to MITRE. This gives the user the choice to choose which dataset to go with based on the perspective of quality or relevance or the overall fitness.

Table 8. Comparison of metrics – Image

Dataset	Business question	Data quality score	Relevance score	Fitness score
Covid-19 Image	Most relevant	96.84%	93.75%	93.75%
Covid CXR	Least relevant	95.72%	84.37%	84.37%

Overall, for our Business Question Mapping Covid-19 Image dataset provides most information. In case of data fitness Covid-19 Image dataset is best fit. Identify covid cases with help of machine learning by processing the images of chest x-rays.

Section 6: Summary

Our framework provides a means to link business questions and tasks to better differentiate data using advanced metadata tagging and indicators for data fitness. By this, manual and iterative task will be reduced that requires domain and technical expertise. Data quality is checked for each dataset in terms of consistency, accuracy, completeness, uniqueness, and timeliness. Precision and relevance are also calculated, which are the basic metrics for data relevance. Business questions are mapped to DLM steps. The relevance of datasets is evaluated and then fitness indicators are generated, which answers the business questions. The results are compared and evaluated for the most fit dataset that provides the client with enhanced information on their business needs which will lay a solid foundation for the future.

Section 7: Future Work

Due to time limitations, we could not identify and detect COVID words from articles or videos. Furthermore, we are unable to automate data tagging using any tools. So, in the future data catalog can be done by using tools like Alation for both text and image datasets.. Also, for Image dataset researchers can integrate EHR data with the COVID-19 image dataset to develop more comprehensive models for predicting COVID-19.

Appendix

Appendix A: Glossary

Term	Definition
LDA	Latent Dirichlet Allocation (LDA) is a popular topic modeling technique to extract topics from a given corpus
CNN	A convolutional neural network (CNN) is a network architecture for deep learning that learns directly from data. CNNs are useful for finding patterns in images to recognize objects, classes, and categories
Momentum AI	Low code/No code platform for machine learning
Confusion Matrix	Shows the performance of a model as a 2x2 table
BERT Model	Language model for sentiment analysis, text summarization, text prediction, text generation, question-answering.
Coherence Score	Measures the degree of semantic similarity between high scoring words in the topic
Zero-shot classification	Supervised learning method where a model is trained on a set of labeled examples but classifies new examples from previously unseen classes.

Appendix B: GitHub Repository

Overview

Our goal is to use advanced data tagging with metadata for data fitness and quality to reduce the manually intensive work in data management life cycle. It ensures the metadata accurately enables discovery, access, and usage of the data for data fitness and quality. Furthermore, we use data analytics and AI/ML techniques to improve this situation.

The primary objective of this project is to better enable users to differentiate datasets that are more likely to meet their business needs. This helps users and organizations to save time and resources to refocus human efforts toward tuning parameters instead of data entry of metadata.

Our process helps users and organizations to incorporate data fitness measures or indicators into DLM processing that consider business questions and tasks by using meta data tags. Data tagging is the process of identifying, categorizing, and labeling data in order to make it easier for a machine learning model to learn. Our solution is to develop a topic modeling algorithm (LDA) that can be used to find topics and thus automate the metadata tagging process. Data tagging can help to improve the quality of data throughout its life cycle. We also aim to find the covid related keywords from business questions along with assessing the data quality and relevance indicators to measure data fitness.

Out of the three text datasets, we are finding out which dataset gives better information about Covid-19. This model gives the user/business enough information to choose the most fit data for their business needs.

- Data Collection: we collected semi-structured and unstructured data from publicly available sources and detected the data types.
- Data Quality Assessment: we measured data quality by assessing four quality dimensions (Accuracy, Completeness, Consistency, Uniqueness).
- Data Cleaning: Data is cleaned by tokenization, removing punctuation, and removing stop words.
- Topic Modeling: we used Latent Dirichlet Allocation (LDA) and it is part of Python's Gensim package.
- Data Relevance Assessment: we calculated coherence score (LDA) and similarity score (BERTopic) to find the most relevant information regarding covid topics or business questions.

On the other hand, we have two image datasets namely covid-19 and COVID CXR image datasets which gives visual information of chest x-rays predicting the possibility of covid. Following back to our business question to predict which dataset best gives covid information we have performed the following steps.

- Data Collection: we collected unstructured data from publicly available sources and detected the data types.
- Data Quality Assessment: we measured data quality by assessing four quality dimensions (Accuracy, Completeness, Consistency, Uniqueness).
- Data Cleaning: Data is cleaned by removing duplicate images, checking consistency, and removing corrupted files.
- Modeling: we used convolutional neural networks (CNN), a machine learning algorithm in momentum AI to predict covid from the chest x-ray images.

We compared the results of quality and relevance metrics about which dataset provides client enhanced information that allows them to easily differentiate datasets that best address their business questions and tasks.

GitHub Repository Link

https://github.com/Skittisu/The-Golden-Hawks

GitHub Repository Contents

The GitHub repository contains the description and python files for data cleaning, data collection, data quality assessment, data relevance assessment, topic modeling, and the final report and presentation slides.

Appendix C: Risks

Sprint 1 Risks

Risk Name	Description	Probability	Impact	Mitigation
Scope Creep	Adding of functionality without addressing the effects on time	Medium	Medium	Determine the project scope.
Accessibility	Data is useful if it can be accessed and used.	Medium	High	Selection based on internal competency.
Enhanced Transparency	Errors, inaccuracies, and bias are common in big data analysis.	Medium	High	Automating processing operations and decision-making processes, including the property factor.
Bad data	Collect every dataset and analyze it later.	High	High	Obtaining data that is irrelevant, up to date and accurate.

The team was able to correctly identify the first three risks. We ran into the final risk of selecting datasets that were unable to answer business questions, which had a significant impact on our project. The lesson we've learned is to ask business questions in advance and select the right data set, then analyze the data first before fully implementing it.

Sprint 2 Risks

Risk Name	Description	Probability	Impact	Mitigation
Unorganized Data	Complicated data and not completely cleaned.	Medium	High	well-planned strategy to collect, store, diversify, eliminate, and optimize data to find meaningful insights
Data Storage and Retention	Big data needs to take a big server area where all the big data is stored, processed, and analyzed.	Medium	High	Take advantage of cloud-based services to store data and make access easy and secure.
Poor Data Quality	Reach poor quality, irrelevant or out-of-date databases that will not help to find something meaningful.	Medium	High	Try to eliminate irrelevant data, and focus on analyzing relevant data to get meaningful insights.

Sprint 2 has three risks. The first is the lack of organization of the data, which is the data is not clean, affecting the results of the analysis, for this reason we need to strategically plan well and clean the data properly. The second is the problem of big data storage, for this reason we should take advantage of cloud-based services to store data. The last is access to data is poor quality and out of date, hence we need to focus on analyzing relevant data to gain meaningful insights.

Sprint 3 Risks

Risk Name	Description	Probability	Impact	Mitigation
Strategy Risk	not fully understanding the capabilities of algorithms and choosing wrong or impossible initiatives.	Medium	High	Understand how algorithms affect people and processes, and clarify clear goals or success metrics

Technical Risk	Team is under-resourced, and skills are not aligned.	Medium	High	Improve team's skills, invest in modern data infrastructure, and adhere to ML best practices.

The team was able to accurately identify the two risks and participate in the action. We often ran into the risk of lack of knowledge and not a full understanding of the capabilities of the algorithm, which greatly affects the outcome of the project and can delay the project. The lesson we learnt was to deeply understand each step and examine the limitations of each algorithm.

Sprint 4 Risks

Risk Name	Description	Probability	Impact	Mitigation
Choosing the Best Model	being able to select a model that closely matches the nature of our data	Medium	Medium	Research in multiple techniques and previous project
Data quality control	Machine learning models are capable of conceiving insights from the data and act accordingly	Medium	High	This behavior demands clean and quality data, else the algorithms will fail to provide the expected accuracy
Hyperparameters	These parameters have a significant role in optimizing the algorithm.	Medium	High	should be configured the default value or tuning Hyperparameter before running the model

Due to our limited knowledge of algorithms, we had problems with proper modeling as well as data quality control which has a big impact on machine learning accuracy. We have considered and worked out how to mitigate this risk by researching several techniques and previous projects and tweaking the hyperparameter before running the model.

Sprint 5 Risks

Risk Name	Description	Probability	Impact	Mitigation
Inefficient delivery	The graph display may not be able to show the user to understand the implementation and results.	Medium	Medium	creating a user interface and data dashboard.

After we get the appropriate algorithm results. We started to find a way to show results. The graph display may not be able to show the user to understand the implementation and results of the work. We have to create user interfaces and dashboards to collect all results on one page. This allows users to have a clearer overview of the project and isolate datasets that are likely to better meet the business needs. Identifying this risk early gives us time to design and implement corrective actions.

Appendix D: Agile Development

- 23/01 - 02/13
- 02/14 - 03/06
- 03/07-03/27
- 03/28 - 04/17
- 04/18 - 05/08

Scrum Methodology

Scrum methodology has worked quite well for our team in terms of collaboration, flexibility, and iterative progress towards the final goal. In the initial stages of the project, we discussed and assigned roles among ourselves as the Product Owner, Scrum Master, and the Developers. We have a sprint planning meeting and then work on the tasks during the sprint. We meet twice/thrice a week to exchange status updates, reviewing the work done and identifying any issues. At the end of each week, we have a presentation, where we find areas of improvement based on the Professor's feedback. With YouTrack as the assigned project management tool, we were able to work together, track tasks and issues, and manage workflows. Since it is customizable, we could create our own workflows, issue types, and custom fields to fit our needs and processes. It has collaborative features with which we could add comments, mentions and notifications. We like that it integrates with Github, since we are creating a GitHub repository for the project. With YouTrack's dashboards and reports, it is quicker to gain insights based on progress and performance.

Sprint 1 Analysis

In Sprint 1, we learned about the metadata tagging domain area. A problem statement and solution space were developed after conducting background research. We also set the project objectives, developing primary user stories and the product vision. We have had meetings twice or thrice a week. Tasks were assigned among the group members and were delivered before the deadlines.

Sprint 2 Analysis

This Sprint is crucial among all the Sprints as this is where most of the data selection and activities like data preprocessing took place. At the initial stages it is understood our focus would be tagging. Structured, unstructured, and semi-structured data were used to prepare our data inventory. We have gathered data in a

variety of formats. The team's main goals for this sprint were to do exploratory data analysis (EDA) on five different datasets and to prepare the data for analysis by transforming both structured and semi-structured data. A schematic diagram is created that outlines the data sources, formats, and data preprocessing stages. Datasets have had their data quality assessments done, and the storage medium, costs, and risks have all been evaluated.

Sprint 3 Analysis

Sprint 3 was full of ideas and thoughts. We came up with business questions for both image and text datasets. To answer the business question, we shared every resource with each other and in every meeting, we went over those resources and how we can incorporate them into our project. We have explored different models for image dataset and have chosen Convolution Neural Network, and for the text dataset Latent Dirichlet Algorithm is used as these would be better models to answer our business questions.

Sprint 4 Analysis

In Sprint 4, we worked on creating visualizations, training the machine learning models, then evaluating and validating them. We also refined our abstract by taking into consideration the changes we made as we progressed in each sprint.

Sprint 5 Analysis

In Sprint 5, we put together our findings, summary of the entire process and identified areas for future work. We also switched to using Momentum AI for image datasets. It was a busy sprint with preparing ourselves for the showcase rehearsals.

Reference

Section 8: Bibliography

[1] TechTarget, "Big Data Management," January 2022. [Online]. Available: https://www.techtarget.com/searchdatamanagement/definition/big-data#:~:text=What%20are%20examples%20of%20big,mobile%20apps%20and%20social%20networks. [Accessed 12 February 2023].

[2] Intellipaat, "Top 10 Top 10 Big Data Applications in Real Life," 10 February 2023. [Online]. Available: https://intellipaat.com/blog/10-big-data-examples-application-of-big-data-in-real-life/#:~:text=level%20of%20traffic.-,Example,be%20set%20for%20every%20trip.. [Accessed 12 February 2023].

[3] M. Whitehorn, "Big Data bites back: How to handle those unwieldy digits," 27 August 2012. [Online]. Available: https://www.theregister.com/2012/08/27/how_did_big_data_get_so_big/. [Accessed 12 February 2023].

[4] Simplilearn, "Challenges of Big Data: Basic Concepts, Case Study, and More," 12 December 2022. [Online]. Available: https://www.simplilearn.com/challenges-of-big-data-article#:~:text=With%20vast%20amounts%20of%20data,be%20stored%20in%20traditional%20databases. [Accessed 12 February 2023].

[5] Tableau, "Data Management: What It Is, Importance, And Challenges," [Online]. Available: https://www.tableau.com/learn/articles/what-is-data-management#:~:text=Data%20management%20is%20the%20practice,the%20vast%20quantities%20of%20data. [Accessed 12 February 2023].

[6] opendatasoft, "What is metadata and why is it as important as the data itself?," 25 August 2016. [Online]. Available: https://www.opendatasoft.com/en/blog/what-is-metadata-and-why-is-it-important-data/. [Accessed 12 February 2023].

[7] University of Texas Libraries, "Metadata Basics," 5 April 2022. [Online]. Available: https://guides.lib.utexas.edu/metadata-basics/key-concepts#:~:text=There%20are%20three%20main%20types,title%2C%20author%2C%20and%20subjects.. [Accessed 2023 February 2023].

[8] University of North Carolina Chapel Hill, "Metadata for Data Management: A Tutorial: Why Do I Need It?," 21 December 2021. [Online]. Available: https://guides.lib.unc.edu/metadata/importance#:~:text=Metadata%20ensure%20that%20data%20are,unless%20textual%20metadata%20are%20available. [Accessed 12 February 2023].

[9] Tibco, "What is Data Discovery?," [Online]. Available: https://www.tibco.com/reference-center/what-is-data-

discovery#:~:text=Data%20discovery%20involves%20the%20collection,framework%20to%20understand%20their%20data. [Accessed 12 February 2023].

[10] A. Morris, "Data Discovery: What Is It & Why Is It Important?," 7 July 2021. [Online]. Available: https://www.netsuite.com/portal/resource/articles/erp/data-discovery.shtml. [Accessed 12 February 2023].

[11] Google Cloud, "What is Data Governance?," [Online]. Available: https://cloud.google.com/learn/what-is-data-governance. [Accessed 12 February 2023].

[12] IBM, "What is data lifecycle management?," [Online]. Available: https://www.ibm.com/topics/data-lifecycle-management. [Accessed 12 February 2023].

[13] M. Andrews, "AI-based auto-tagging of content: what you need to know," 8 August 2022. [Online]. Available: https://kontent.ai/blog/ai-based-auto-tagging-of-content-what-you-need-to-know/. [Accessed 12 February 2023].

[14] U.S. Securities and Exchange Commission, "Data Tagging," [Online]. Available: https://www.investor.gov/introduction-investing/investing-basics/glossary/data-tagging. [Accessed 12 February 2023].

[15] V. Mahn-DiNicola, "Six Dimensions of Data Fitness," 25 January 2019. [Online]. Available: https://blog.medisolv.com/articles/six-dimensions-of-data-fitness. [Accessed 12 February 2023].

[16] M. Suer, "What Is Data Quality and Why Is It Important?," 5 August 2021. [Online]. Available: https://www.alation.com/blog/what-is-data-quality-why-is-it-important/#:~:text=Data%20quality%20is%20defined%20as,used%20for%20a%20given%20purpose. [Accessed 12 February 2023].

[17] R. L. Sarfin, "5 Characteristics of Data Quality," 2 November 2022. [Online]. Available: https://www.precisely.com/blog/data-quality/5-characteristics-of-data-quality#:~:text=Read-,Relevance,for%20the%20sake%20of%20it. [Accessed 12 February 2023].

[18] T. Regan, "The Importance of Data Quality for Business," 22 August 2022. [Online]. Available: https://www.reworked.co/information-management/the-importance-of-data-quality-for-business/. [Accessed 12 February 2023].

[19] R. Atarashi, J. Kishigami and S. Sugimoto, "Metadata and new challenges," 27 January 2003. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/1210192. [Accessed 12 February 2023].

[20] M. M. I. M. A. M. Thomas Margaritopoulos, "A Conceptual Framework for Metadata Quality Assessment," 2008.

[21] F. G. Wenfei Fan, "Foundations of Data Quality Management," August 2012. [Online]. Available: https://dl.acm.org/doi/10.5555/2371176. [Accessed 12 February 2023].

[22] C. C. C. F. A. M. Carlo Batini, "Methodologies for data quality assessment and improvement," *ACM Digital Library,* vol. 41, no. 3, pp. 1-52, 2009.

[23] G. Preda, "BBC News," 19 February 2023. [Online]. Available: https://www.kaggle.com/datasets/gpreda/bbc-news. [Accessed 20 February 2023].

[24] Shawn, "Flickr30," 16 January 2023. [Online]. Available: https://www.kaggle.com/datasets/eeshawn/flickr30k?select=captions.txt. [Accessed 20 February 2023].

[25] C. Rafael, "Tweets Blogs News - Swiftkey Dataset 4million," 20 February 2023. [Online]. Available: https://www.kaggle.com/datasets/crmercado/tweets-blogs-news-swiftkey-dataset-4million. [Accessed 20 February 2023].

[26] H. Dorney, "8 useful insights you can learn from Twitter analytics," Twitter, [Online]. Available: https://business.twitter.com/en/blog/7-useful-insights-twitter-analytics.html. [Accessed 20 February 2023].

[27] A. York, "How to analyze Twitter data," 10 December 2020. [Online]. Available: https://sproutsocial.com/insights/twitter-data/. [Accessed 20 February 2023].

[28] J. Leskovec, "Flickr image relationships," 2012. [Online]. Available: https://snap.stanford.edu/data/web-flickr.html#:~:text=Dataset%20information,images%20taken%20by%20friends%2C%20etc. [Accessed 20 February 2023].

This page intentionally left blank.