

# FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

## Declaration

Plagiarism is defined as “the unacknowledged use, as one’s own work, of work of another person, whether or not such work has been published” (Regulations Governing Conduct at Examinations, 1997, Regulation 1 (viii), University of Malta).

I / We\*, the undersigned, declare that the [assignment / Assigned Practical Task report / Final Year Project report] submitted is my / our\* work, except where acknowledged and referenced.


I / We\* understand that the penalties for making a false declaration may include, but are not limited to, loss of marks; cancellation of examination results; enforced suspension of studies; or expulsion from the degree programme.

Work submitted without this signed declaration will not be corrected, and will be given zero marks.

\* Delete as appropriate.

(N.B. If the assignment is meant to be submitted anonymously, please sign this form and submit it to the Departmental Officer separately from the assignment).

Kian Parnis  
Student Name

  
Signature

\_\_\_\_\_  
Student Name

\_\_\_\_\_  
Signature

\_\_\_\_\_  
Student Name

\_\_\_\_\_  
Signature

\_\_\_\_\_  
Student Name

\_\_\_\_\_  
Signature

ARI2201  
Course Code

ARI2201-Artifact-and-Report-Submission-Kian-Parnis-0107601L  
Title of work submitted

6/13/2022  
Date

# Intelligent Information Dashboard

Kian Parnis

Department of Artificial Intelligence  
kian.parnis.20@um.edu.mt

Alexiei Dingli

Supervisor  
alexiei.dingli@um.edu.mt

## ABSTRACT

Analyzing Big Data is the process of handling large quantities of complex and large data sets by cleaning, transforming, and building a model which aims to capture useful information from patterns and trends, to better improve on making informed decisions. Social Media Platforms such as Facebook, Twitter and Social news outlets have a constant stream of information with new emerging topics and trends constantly on the horizon. Natural Language Processing (NLP) tools makes it possible to connect to this stream of Big Data and enables the processes such as data analysis to give a better understanding on how it could be understood. Topic Modelling (TM) aims to grab large collections of documents and formulate “topics” which documents can be sorted in. Sentiment Analyses (SA) tackles the understanding of human language to determine if complex sentences are in a range of three distinct categories, positive, negative, or neutral to better show the underlying emotion behind text. Finally, Named-entity Recognition (NER) classifies data by finding and categorizing text into key features, such as individual’s names, brands, and locations. By crafting a solution that makes use of these various methods, capturing and analyzing large chunks of data into useful knowledge from various media outlets are enabled.

## Keywords

Natural language processing; Topic modelling; Sentiment Analyses; Name Entity Recognition

## 1. INTRODUCTION

### 1.1 Aims and Objectives

The aim of this IAPT is make use of these natural processing tools to design a dashboard which allows users to harvest large chunks of data from various amounts of sources, being able synthesize data into different “topics” and constructing labels to create a better cohesive view of what topics may be emerging.

The following are the exact objectives defined for this IAPT:

#### Harvesting Phase

1. Create a scraper which regularly downloads top pages from Google based upon specific search terms (which are provided in the settings)
2. Create a scraper capable of harvesting information from other websites, online newspapers, and others (which are provided in the settings).
3. Create a scraper capable of monitoring tweets of influential people on twitter, or groups on Facebook, LinkedIn, etc. All these will be set in the settings.

#### Analysis

1. Create a web interface which gives statistics on important topics (Identify the most important topics over time, look for emerging topics, etc.)

2. Allows the user to create a monitoring alert which creates a tab on the webpage summarizing that topic, and when a particular topic gets prominence, an email is sent to the use which gives a summary of the findings.

### 1.2 Summary of Solution Developed

Using various Application Programming interfaces (API) a user can request to receive specified amounts of data from specific outlets, such as Twitter with the use of the Tweepy API, Facebook using the facebook\_scraper API as well as various APIs designed to retrieve data from google and various news outlets (Search and Article APIs). This data would then be present to the user in its entirety separated in rows and columns of data and links, linking back to the original material. Once the user is satisfied, they can opt to process this data which goes through the following techniques:

Sentiment Analyses (SA), the openly available Valence Aware Dictionary and Sentiment Reasoner (VADER) model provided by the Natural Language Toolkit (NLTK) [1] to label the individual documents.

Topic Modelling (TA), sklearn’s latent dirichlet allocation (LDA) model was used to distinguish each document into its own distinct category “topic” [2].

Named Entity Recognition (NER), spacy’s NER’s were used to identify entity names (People, Places, Brands etc.) from each document within a topic [3].

The findings are then reported back to the user separated into two distinct tabs, the first being the results of the NER sorted by rows and columns of Topics followed with the number of times particular entities are mentioned. The second tab consists of each topic organized alphabetically with the individual documents split into three columns: a summary of the text collected (Using Articles summary functionality), the SA score (calculated on the data before summarization) and finally a reference back to the original source.

At the end of this tab the user is given an option to insert their email address to receive an email which contains the final findings.

### 1.3 Instructions for setup of artifact

When setting up the artifact solution to work with flask, a python environment was created which contains all the packages used.

Table 1. List of packages and their respective versions

Package	Version
Flask	2.1.1
Flask-Mail	0.9.1
Google	3.0.0
Beautifulsoup4	4.11.1
Facebook-scraper	0.2.55
Newspaper3k	0.2.8
NLTK	3.7

Numpy	1.21.6
Scikit-learn	1.0.2
Scapy	1.7.3
Tweepy	4.8.0
Requests	2.27.1
Regex	2022.3.2
Pillow	9.1.0
Python	3.7.9

This venv is labelled as 'IAPT\_env' and to start hosting the web service one needs to first load the bat file within the VENV using "IAPT\project\_env\Scripts\activate.bat" from an open command line, once activated one would then need to redirect via the command line to the 'Flask Server' folder and write these commands in sequence "set FLASK\_APP=Flask.py" followed by "python -m flask run" the command line will now present a link and one can copy this address given and paste it in google to load the artifact.

## 2. LITERATURE REVIEW

Throughout this literature review, a more in-depth lens will be taken on the core AI techniques which have been utilized in the data processing phase. This pertains to, Sentiment Analyses, Topic Modelling, TF IDF model and finally Named-entity Recognition.

### 2.1 Sentiment Analyses

Sentiment Analyses also known as opinion mining is a subfield of natural language processing, it's an active research field that has recently emerged and it aims to extract people's opinions, feelings, sentiments from data in the form of text by utilizing natural language processing methods [4]. Sentiment Analyses covers a variety of classifications such as splitting classifying text into categories like "positive" or "negative" or "neutral". Kharde and Sonawane in [5] break down opinions and sentiments as being interchangeable between:

**Opinion:** where different experts in fields may have different opinions

**View:** one's own subjective opinion

**Belief:** one's faith, trust and confidence in something or someone

**Sentiment:** a feeling being represented by one's opinion

They [5] also processed to give the following example to better represent these terminologies:

**Full sentence:** The story of the movie was weak and boring.

**Object within sentence:** movie

**Feature within sentence:** story

**Opinion within sentence:** weak, boring

**Final Polarity attributed:** negative

Such an example gives a final polarity of negative towards a particular review of a movie, Sentiment Analyses can also be used to classify sentiments for products, services organizations etc. As mentioned prior, sentiment analyses are a very active area of research thus, various approaches have been taken by different research such as [6] which aimed to classify twitter tweets by

utilizing twitter emoticons and utilizing the naïve bayes method. The widely available VADER model was chosen which deals with polarities of negative/positive and strength of one's sentiment [1]. This model was chosen due to the general classification which is achieved allowing one to accurately classify different social media text, as well as its effectiveness compared to various approaches taken such as: Support Vector Machines and Naïve Bayes which resulting in its best accuracy when classifying tweets at an F1 score 0.96 vs SVM's 0.83 and NB's (tweets) 0.84 [1].

### 2.2 Topic Modelling

Topic modelling is an unsupervised machine learning approach technique that examines a collection of documents, discovers word and phrase patterns, and automatically clusters word groupings and related phrases that best describes a set of text documents. To date LDA (Latent Dirichlet Allocation) has been one of the most popular techniques used in a wide variety of applications for topic modelling with the original paper [7] being cited approximately 42,892 times (as of June 9th, 2020). LDA is a generative probabilistic model that generates topics based on word frequency from a set of documents. During data processing punctuation and stop words are excluded from a particular topic [8] this results in a vector of words which contain a high amount of topical content that can be used for topic categorization. Figure 2.1 shows how LDA uses this notion for modelling.

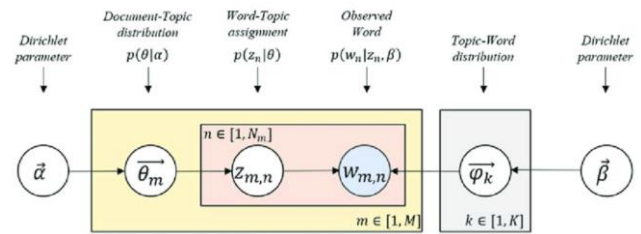


Figure 2.1 is a Graphical Model of Latent Dirichlet Allocation (LDA) [9]

Table 2: Breakdown of the notations featured in figure 2.1

Symbol	Meaning
$\theta_m$	Indicates the distribution a particular document $m$ may have relative to a set of documents example: document A might have a distribution of 0.5 to science, 0 to Sport and 0.5 to Religion.
$m$	Represents the total number of documents within a corpus.
$n$	Represents the total number of words found within a document.
$k$	$k$ in Topic-word distribution highlights the number of topics we would want to classify such as science, sport, religion as $K=3$ .

$\alpha, \beta$	The Dirichlet parameters alpha and beta represent the document-topic distribution and the topic word distribution respectfully. A high alpha would indicate that each document is likely to contain a mixture of most of the topics and not just one or two. Conversely a low alpha indicates that each document will likely contain just a few of the topics. A high beta indicates that each topic will contain a mixture of most of the words while a low Beta may show that each topic may contain a mixture of just a few words.
$z_{m,n}$	Denoted as the topic for the n-th word in document m.
$w_{m,n}$	The specified word.

### 2.3 TF IDF

Latent Dirichlet Allocation utilizes TF IDF which transforms cleaned text into numerical representation. TF IDF stands for term frequency – inverse document frequency, this can be split into two categories i.e., the calculation of TF followed by the calculation of IDF. TF is calculated based on the frequency a particular word ‘t’ appears in document ‘d’ TF(t,d) the following is an example: given word “Technology” which appears exactly 5 times in a document containing 2500 words, the Term frequency would be 5/2500 which results as 0.002. IDF on another hand prioritizes terms with low document frequencies this uses df which is the document frequency i.e., the number of documents in which ‘d’ appears such that  $idf = \log\left(\frac{N}{df(t)}\right)$  where n is the number of documents present. The result would be the multiplicative result of Tf x Idf [10]. Tf idf gets mostly limited as noted in [11] when slight changes in words occur for example, TF IDF would treat words such as “play and “playing” as two distinct words would result in undesired results. A solution to this, which is used within the IAPT is by conducting a stemming process which would take our two words and craft the stem of both to the word “played”.

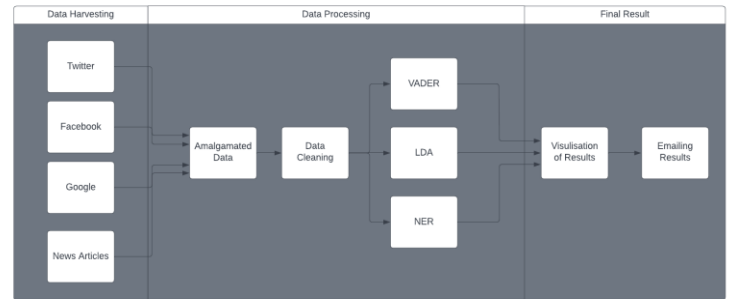
### 2.4 Named-Entity Recognition

Named-entity recognition also known as (named) entity identification is the process of classifying words and/or expressions into different Named Entity [NE] Categories [12]. Spacy’s en\_core\_web\_sm model, which is the model used by the system is a neural network model, which is commercially opensource and is easy to deploy [13]. This model is used primarily for the English language and performs efficiently and accurately with a Named Entity F-score of 0.85, recall score of 0.84 and finally, a precision score of 0.92 [14].

## 3. METHODS AND MATERIALS

In Figure 3.1, the full process achieved by this IAPT is showcased as a chart of the full breakdown of each step taken from harvesting data, to Data Processing and showcasing the final results.

**Figure 3.1: Flowchart of the Method of implementation**



### 3.1 Data Harvesting

When harvesting Data, a user can specify specific information to receive and the quantity of how much information they wish to receive. The following is a breakdown of each of the four different data collectors and how they go about retrieving data.

#### Twitter API:

The user enters the specific information they would like to receive prefixed by hashtag examples: #technology, #IT #Devops etc. and the exact number of tweets they would like to receive. The full text of the top x most recent tweets in their entirety are then received followed by the links to the original tweets. When acquiring tweets, retweets are filtered out to avoid the possibility of receiving duplicate tweets and the tweets written in English are exclusively harvested.

#### Facebook API:

Information is received uniquely from Facebook groups; thus, the user would need to specify the name of the group (Private groups cannot be accessed). The API [15] functions by crawling between different pages within the group (First two pages might not yield a result) so an option for a specific page amount is given.

#### Google API:

The user can request any search term from the internet and depending on the size requested, the exact amount of googles top search results URLs are first received, this is then synchronized with the news article API to parse and extract all the relevant text found within the URLs themselves.

#### News Articles API:

Newspaper [16] allows several features for extracting information from websites, this can be used both for article extracting as well as general website extracting, the user would need to pass in a specific URL to an article, then the API proceeds to attempt to download every article found within the page and extract each articles data.

### 3.2 Data Cleaning

Once the data is collected in a raw format, cleaning would need to take place to remove unwanted information that might lower the quality of analysis. The following is the process the data goes through for cleaning:

- Any links found within text are filtered out.
- #, @, \$, ! symbols are removed.

- White spaces are removed.
- All the data is set to lower case
- Stop words are removed
- Individual words are broken down to their stem

For stop word removal, NLTK's stop word corpus library of English stop words is used. The following is a list of the full contents of stop words found in the NLTK corpus [17]. NLTK's Porter Stemmer is also used to break down each word to its root/base words to minimize the variants of words which is based on the following algorithm [18].

### 3.3 Sentimental Analyses

The VADER model provided by NLTK was used to go through the individual documents collected and give a ranking of Positivity / Neutrality / Negativity. These scores are assessed out of a total of a 100% split among these categories (Each are kept at two decimal places).

### 3.4 Topic Modelling

When it came to topic modelling a specific number of topics were chosen beforehand, these are Topics A-H. This has been tested to give the best generalization of topics as the dataset groups grow larger. Before the Latent Dirichlet Allocation model could be used, the cleaned data had to be transformed into the Vector Space Model: TF IDF and after fitting this with LDA. Information such as links and SA score are retained, and a summary of the original text is achieved with Newspaper3k to give users better readability. Finally, after the model has been fit, a sorting pass was taken on the topics to maintain an alphabetical order of documents.

### 3.5 Named Entity Recognition

The NER was used to count the number of times specific entities such as People, Places and Brands appear within all the documents present within a particular Topic. For each topic a sorted count of was then displayed to the user along with the topics themselves. A choice was taken to display the top five most mentioned entities within each document to keep a balance of informative while not presenting too much (if the dataset is smaller).

### 3.6 Visualization

The Flask framework [19] was used after all the prior individual components were finalized within a Jupiter Notebook as a backend to house all the major workings of the solution, this would allow the user to make requests on the frontend (html), process this request and achieve a result and then finally return the result back to the frontend where the results would be shown in a presentable way. The following is a breakdown of all the individual aspects within the final dashboard.

#### First access:

Upon first accessing the site one tab is shown that allows the user to enter specific information for each of the four retrieval methods {Twitter, Facebook, Article, Google} a user can simply choose to allocate more input boxes to add more websites, hashtags, groups etc. and upon doing so the user is able to press 'add data' which starts the harvesting phase.

#### Data Collection:

When the data is harvested a new tab appears below the dashboard which presents all the data retrieved by the harvesters, these are in the form of data | source link. The user is given a couple of options to proceed: they can delete the data and attempt to retrieve other types of data, they can go back and choose to continue to harvest

more data, and this is added to the current data kept and finally, once they decide that the data is sufficient, they can start the processing stage. This will then proceed to add two new tabs which showcase the different results present. Fig 3.6.1 illustrates an example of the output of a sample data harvest.

#### NER results:

The tab labelled 'Mentioned Brands/People in particular topics' presents all the topics A-H alongside the results of the named entity recognizer. As already stated, these are the results from iterating over each document within a topic and taking the top five most mentioned entities fig 3.6.2 showcases a sample result of what can be achieved.

#### LDA and VADER results:

The final tab 'Processed Results' presents all the topics A-H with the formulated keywords of the topic with the documents of their respective categorization. Each document is presented in the form {Text Summarization | Sentiment Score | Reference}. Finally, below the last topic the user can choose to enter an email address and all the processed data within this tab will then be sent to that respective email address by the system. This processed is achieved with the use of Flask's mail feature [20]. Fig 3.6.3 and Fig 3.6.4 showcase a sample output as well as the mailing section of this tab.

Figure 3.6.1: A sample of the output data harvest

Current data collection:

luxurytravel aviation trip business luxury businessaviation future technology turbo	
Market analyst from ABI recently projected that between 2021 and 2026, the number of IoT connections will reach 23.6 billion, each one representing an opportunity to leverage AI, machine learning and TinyML, at technology BigData ML.	https://twitter.com/tweet/statuses/153372027548411264
10 Great FOSS Linux Games You Never Knew Existed. technology	https://twitter.com/tweet/statuses/153372003517230563
Happy Monday 🌞🌞🌞 "They fear love because it creates a world they can't control." George Orwell. 1984 a1thingsdata governance a1thingsdata DataAnalytics data protection STEM technology womenindata womenintech WomeninSTEM	https://twitter.com/tweet/statuses/1533719614598567636
BMW's M Hybrid V8 LMH Race Car Debuts Red Livery for Testing - ONETechne technology womenintech	https://twitter.com/tweet/statuses/1533719883594203137
We're living in a post-breach era! and no company is safe cybersecurity ciso technology via vintedhawk	https://twitter.com/tweet/statuses/1533719844104855563
Welcoming onboard our Silver Partner Adobe for the FEModernBI-SI 2022. Register Now - FEModernBI-SI Summit Ranking RFSI Technology Innovation RFSI Summit AdobeSoftware AI data cloud datamanagement data science artificialintelligence	https://twitter.com/tweet/statuses/153371976966472632
Construction Technology Festival 2022 is back with its 5th edition - Construction and Civil Engineering News digitalwins technology trends	https://twitter.com/tweet/statuses/1533719729742835712
How did our ancestors value honey? — They purchased honey with their death tax.	

Retrieve Data Process Data

Figure 3.6.2 Sample result of Named Entity Recognition.

Mentioned Brands/People in particular topics:

(Topic's followed with the amount of times particular entities are mentioned)

Topic: A [agriculture] industry size futureoutag scope growth market	[2020R, 6], (Philippines, 2), (Fridge, 2), (Whole Foods, 2), (one day, 1)
Topic: B [technology] cybersecurity infocast hack news technologynew report	[HACKER, 1], (100DaysOfCode, 1), (DataSecurity CyberSec Hackers, 1), (Mediteranean, 1), (menorquin34 Host, 1)
Topic: C [future] technology news mental impact technologynew implement	[Russian, 3], (Seattle, 2), (the Ray Area, 2), (Allen Institute, 2), (May, 2)
Topic: D [technology] take construct great foldable phone work	[Traf, 4], (Blue Origin, 3), (10, 2), (Pakistan, 2), (ICT, 2)
Topic: E [technology] tech work year news last shoot	[Yesterday, 4], (Nature Trust, 3), (Valletta, 2), (Texas, 2), (Tulsa, 2)
Topic: F [technology] futur robot news live agricultural hack	[El Sabon, 3], (1968, 2), (1970, 2), (1948, 2), (Omair, 1)
Topic: G [technology] europ cybersecurity fund health men podcast	[Europe, 4], (Mas Zabor, 4), (Mala, 2), (pakistani, 1), (App, 1)
Topic: H [technology] cybersecurity technologynew infocast hack news dataset	[Today, 3], (Miami, 3), (milions, 2), (Ukrainian, 2), (the 100 day, 1)

**Figure 3.6.3 Sample result of Topic modelling Sentimental Analysis of the first two topics.**

## Processed Results:

Topic: A [agricultur' 'industri' 'size' 'futureofag' 'scope' 'growth' 'market']		
hybrid course) For more information and to book, visit: <a href="#">subsystems</a> learninganddevelopment hybridising subsystems engineering technology	Positive: 0.0% Negative: 0.0% Neutral: 100.0%	<a href="https://twitter.com/twitters/statuses/1533720574035472384">https://twitter.com/twitters/statuses/1533720574035472384</a>
Happy Monday! 🌞🌱🌿 "They fear 'love' because it creates a world they can't control." George Orwell, 1984 <a href="#">ethingsdata</a> governance <a href="#">ethingsdata</a> Data/Analytics <a href="#">dataprotection</a> STEM technology <a href="#">womenindata</a> <a href="#">womenintech</a> <a href="#">WomenInSTEM</a>	Positive: 29.2% Negative: 9.4% Neutral: 61.4%	<a href="https://twitter.com/twitters/statuses/1533719614580507636">https://twitter.com/twitters/statuses/1533719614580507636</a>
y: Linolenic Acid Makes Old Blood New Again news technology <a href="#">TechnologyNews</a> <a href="#">infosec</a> cybersecurity hacking	Positive: 0.0% Negative: 0.0% Neutral: 100.0%	<a href="https://twitter.com/twitters/statuses/1533719259381121024">https://twitter.com/twitters/statuses/1533719259381121024</a>
Source code for 4th democratic production "Elevated" released (2016) news technology	Positive: 0.0% Negative: 0.0% Neutral: 100.0%	<a href="https://twitter.com/twitters/statuses/153371920658080081">https://twitter.com/twitters/statuses/153371920658080081</a>

Topic: B [ 'technology' 'cybersecur' 'infosec' 'hack' 'news' 'technologynew' 'appli']		
Text Summary	Text Score	Reference
Social media technical issues and assistances contact HACKER TEL is Machine learning python IoT 100DaysOfCode programming infosec cybersecurity pentesting esxp informationsecurity hacking	Positive: 0.0% Negative: 0.0% Neutral: 100.0%	<a href="https://twitter.com/twitters/statuses/1533721268530185216">https://twitter.com/twitters/statuses/1533721268530185216</a>

**Figure 3.6.4 Showcasing the mailing section of Processed Results tab.**

Hopara's India rebrands to consolidate US\$78 billion Health opportunity healthtech innovation health science technology Covid19 Vaccines pharmaceutical manufacturing	Positive: 19.7% Negative: 0.0% Neutral: 80.3%	<a href="https://twitter.com/twitters/statuses/1533721190761734144">https://twitter.com/twitters/statuses/1533721190761734144</a>
--	---	---

Topic: H [ 'technology' 'cybersecur' 'technologynew' 'infosec' 'hack' 'news' 'datasci']		
Text Summary	Text Score	Reference
. GlobalCrisis. Very true it's so nice we have already these technologies in our power today to save millions of people from hunger! But only when we voiced the truth Together hunger technology	Positive: 29.4% Negative: 5.0% Neutral: 65.5%	<a href="https://twitter.com/twitters/statuses/1533721186051448833">https://twitter.com/twitters/statuses/1533721186051448833</a>
infographic: Here are the 100 days of the DataScience Challenge Via inglipuoi <a href="#">datascience</a> <a href="#">machinelearning</a> <a href="#">python</a> <a href="#">artificialintelligence</a> <a href="#">ai</a> <a href="#">data</a> <a href="#">datasci</a> <a href="#">analytics</a> <a href="#">bigdata</a> <a href="#">programming</a> <a href="#">coding</a> <a href="#">datascientist</a> <a href="#">technology</a> <a href="#">deeplearning</a>	Positive: 5.1% Negative: 0.0% Neutral: 94.9%	<a href="https://twitter.com/twitters/statuses/153372108288945664">https://twitter.com/twitters/statuses/153372108288945664</a>
First NFT insider trading case in US involves former OpenSea employee with knowledge of homepage placement Web3 Technology NFT Community	Positive: 8.8% Negative: 0.0% Neutral: 91.2%	<a href="https://twitter.com/twitters/statuses/153372095572738051">https://twitter.com/twitters/statuses/153372095572738051</a>

## 4. EVALUATION

### 4.1 Data Harvesting

When conducting testing related to the retrieval of data, checks where done that 1) The quantity of data is met and 2) the data is accurately scraped. The latter involved cross checking the data that has been retrieved with the links that are kept by the system, the following is an example of how cross checking was conducted.

**Figure 4.1.1 Showcasing a singular tweet scraped from phrase #futurist**

Full interview with @LeaHogg on technology trends shaping the future! <https://it.coxybyHUB4K#thefuture#futuristic#futurist#virtualreality> <https://twitter.com/twitters/statuses/1535128970416758784>

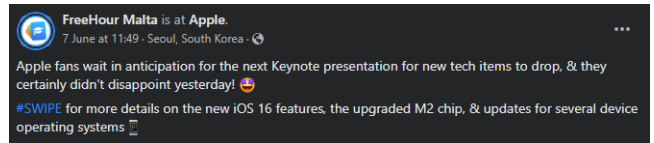
**Figure 4.1.2 cross referencing tweet (Fig 4.1.1) back to source**



**Figure 4.1.3 Showcasing a singular Facebook post scraped from group 'freehourmalta'**

Apple fans wait in anticipation for the next Keynote presentation for new tech items to drop, & they certainly didn't disappoint yesterday! #SWIPE for more details on the new iOS 16 features, the upgraded M2 chip, & updates for several device operating systems <https://facebook.com/freehourmalta/posts/2216111301672787>

**Figure 4.1.4 cross referencing Facebook post (Fig 4.1.3) back to source**



**Figure 4.1.5 Showcasing a singular result scraped from google with term 'IOS 16'**

Wallet Adds Apple Pay Later, Order Tracking, and Other FeaturesApple Pay Later provides users in the US with a seamless and secure way to split the cost of an Apple Pay purchase into four equal payments spread over six weeks, with zero interest and no fees of any kind.3 Built into Apple Wallet and designed with users' financial health in mind, Apple Pay Later makes it easy to view, track, and repay Apple Pay Later payments within Wallet. Users can apply for Apple Pay Later when they are checking out with Apple Pay, or in Wallet. Apple Pay Later is available everywhere Apple Pay is accepted online or in-app, using the Mastercard network.4 Additionally, with Apple Pay Order Tracking, users can receive detailed receipts and order tracking information in Wallet for Apple Pay purchases with participating merchants. Today, Apple leads the world in innovation with iPhone, iPad, Mac, Apple Watch, and Apple TV. Apple's five software platforms — iOS, iPadOS, macOS, watchOS, and tvOS — provide seamless experiences across all Apple devices and empower people with breakthrough services including the App Store, Apple Music, Apple Pay, and iCloud.	<a href="https://www.apple.com/newsroom/2022/06/apple-unveils-new-ways-to-share-and-communicate-in-ios-16/">https://www.apple.com/newsroom/2022/06/apple-unveils-new-ways-to-share-and-communicate-in-ios-16/</a>
--	---

**Figure 4.1.3 cross referencing Google result (Fig 4.1.5) back to source**



### 4.2 Semantic Analysis

As mentioned prior, VADER was used for classifying data into separate classifications, in [1] a metric is provided for the three-class accuracy (F1 scores) for VADER as well as other techniques.



**Figure 4.2 F1 scores for VADER relative to other metrics from [1]**

	3-Class Classification Accuracy (F1 scores)			
	Test Sets			
	Tweets	Movie	Amazon	NYT
VADER	<b>0.96</b>	0.61	<b>0.63</b>	<b>0.55</b>
NB (tweets)	0.84	0.53	0.53	0.42
ME (tweets)	0.83	0.56	0.58	0.45
SVM-C (tweets)	0.83	0.56	0.55	0.46
SVM-R (tweets)	0.65	0.49	0.51	0.46
NB (movie)	0.56	<b>0.75</b>	0.49	0.44
ME (movie)	0.56	<b>0.75</b>	0.51	0.45
NB (amazon)	0.69	0.55	0.61	0.48
ME (amazon)	0.67	0.55	0.60	0.43
SVM-C (amazon)	0.64	0.55	0.58	0.42
SVM-R (amazon)	0.54	0.49	0.48	0.44
NB (nyt)	0.59	0.56	0.51	0.49
ME (nyt)	0.58	0.55	0.51	0.50

This Table was considered when evaluating the implementation of VADER for the system. As shown in shown in Fig 4.3, in the first row we have words that are considered neutral ex: {Daily, Futurist, we} words only a few words that represent positivity ex {Inspiration, Solutions} and finally words such as {whining, whiners, without}. This shows a valid accuracy by VADER and is considered a valid score. With this metric in mind a sample of tweets, posts, articles etc. were taken and evaluated table 4.1 showcases this result.

**Figure 4.3 Two sample tweets and their Polarity score**

Text Summary	Text Score
Daily Inspiration: "Stop whining Start doing" - Futurist Jim Carroll We are surrounded by whiners without solutions.	Positive: 15.7% Negative: 33.4% Neutral: 50.9%
Supporting Users with HCD principals Futurist IoT BlockChain Agile DevOps	Positive: 24.4% Negative: 0.0% Neutral: 75.6%

**Table 4.1 Accuracy received by sampling 30 results from each category**

	TWITTE R	FACEBOO K	GOOGL E	ARTICL ES
VADER	<b>0.93</b>	<b>0.66</b>	<b>0.53</b>	<b>0.53</b>

From Table 4.1 it is shown that VADER managed to maintain a similar score to that of the F1 rankings in [1], This result suggests that VADER tends to be most accurate when dealing with texts of minor length, this is due to amount of neutral text that is present the longer the text is, consider Figure 4.4, the text received from scrapping a particular Facebook page results in text that deals with theft by VADERs accuracy dwindles due to the amount of natural text present with the text.

**Figure 4.4 Sample Post and Polarity given**

3 Romanians have been sentences to 54 months in prison & fined €10,000 each after admitting to stealing money from passengers while riding on buses These 3 men - aged 40, 41, & 55 - are part of a larger international group of criminals involved in pickpocketing. They were accused that in June & July, they stole from 5 passengers on buses, committed money laundering, associated themselves with others to commit a crime & possessed stolen objects. After their stealing spree, the Romanians left Malta. It was later discovered that they had been committing the crimes in various European countries for 11 years. The Maltese Police were then informed that the men were returning to Malta on a flight from Budapest. & arrested them.	Positive: 4.4% Negative: 18.8% Neutral: 76.8%
--	--

### 4.3 Latent Dirichlet Allocation

As already mentioned, LDA was used to split documents into Topics A-H, evaluating these topics was done by generating each topic and verifying if documents are being accurately categorized under a respective topic. In general topics where correctly appointed giving a satisfiable result of how the texts where being split and categorized across the different categories.

### 4.3 Named Entity Recognition

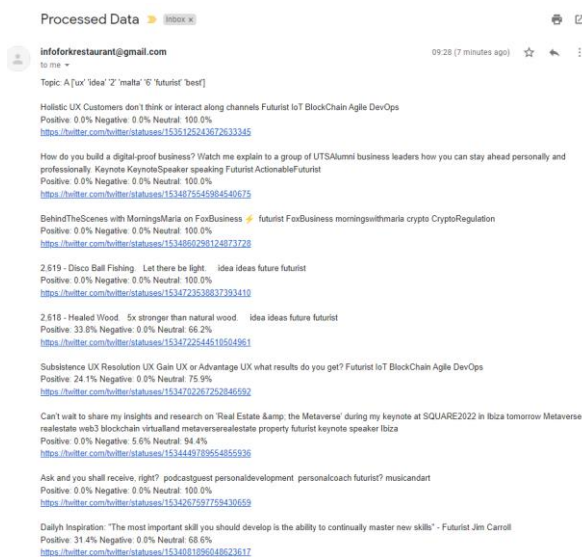
Similarly, to how LDA was evaluated, NER results where cross referenced back to each document present within a particular topic, verifying if the number of entities found matched what is being shown by the topics tab. This results in an accurate measurement in the number of times a particular NER was found throughout a particular document.

### 4.5 Emailing results

To properly verify that emails are being sent as intended to the user upon submission, a handful of emails were sent, and each email was inspected to check that the valid format is kept which is:

- 1) Topic
- 2) Documents
  - 2.1) Document Summary
  - 2.2) Text polarity ranking
  - 2.3) link to source

**Figure 4.5 Sample email received by the user (full email omitted)**



## 4.6 Further evaluation considered

Additional evaluation was taken during the data cleaning phase, this involved sampling text and considering what text might be redundant such as https link which lowered the quality of ai techniques, and these were coded out from the system. Furthermore, tests were done on the dashboard itself to validate a proper user experience would be met, this involved testing the website on different dimensions for usage on a handheld device and verifying that all the inputs a user could take worked as intended.

## 5. CONCLUSIONS

### 5.1 Achieved Aims and Objectives

Reiterating the aim of this IAPT, the goal was to create a dashboard that fulfils a purpose of retrieving emerging data from different social sites, amalgamating this information together and finally using different techniques to make sense and categorize this data. This was achieved with preprocessing techniques that allowed the cleaning of this data while retaining all relevant information suited for processing. The processing stage gave a satisfiable result with the use of the LDA and VADER models as well as NER which are presented in a visualized manner.

### 5.2 Limitations

The most notable limitation was that of time taken for data retrieval, API's such as twitters tweepy offers a direct link for retrieval of data but required a tedious process of admissions so that one could be able to access these tools. For different social media sites such as Facebook an undirect link was established using a crawling method found within the packages used which led to good results but requiring a longer wait time for the full retrieval of data.

### 5.3 Future Work

To expand this dashboard further in the future one can investigate integrating more retrieval APIs to collect vaster amounts of data from other different sources of information, such using Reddit's API to retrieve trending information from the flowing stream of information found within this website. Linked in is another network that could be integrated to even retrieve information about professionals' beings suggested when collecting entities via the NER that's already established.

### 5.4 Final Remarks

The most time-consuming process of this IAPT was researching and evaluating the different small pieces that would in the end together formulate an informative result. The knowledge acquired from testing and researching these different methodologies and what is currently available provided a deeper understanding into Big Data and how different natural language processing techniques attempt to breakdown and formulate an informative lens in acquiring what relevant trends might be emerging and the importance of these trends.

## 6. REFERENCES

- [1] "GitHub - cjhutto/vaderSentiment: VADER Sentiment Analysis. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media and works well on texts from other domains.", GitHub, 2022. [Online]. Available: <https://github.com/cjhutto/vaderSentiment>. [Accessed: 12- Jun- 2022]
- [2] "GitHub - scikit-learn/scikit-learn: scikit-learn: machine learning in Python", GitHub, 2022. [Online]. Available: <https://github.com/scikit-learn/scikit-learn>. [Accessed: 12- Jun- 2022]
- [3] Spacy.io, 2022. [Online]. Available: <https://spacy.io/api/entityrecognizer>. [Accessed: 12- Jun- 2022]
- [4] Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216-225. Retrieved from <https://ojs.aaai.org/index.php/ICWSM/article/view/14550>
- [5] V. A. and S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques", International Journal of Computer Applications, vol. 139, no. 11, pp. 5-15, 2016 [Online]. Available: <https://arxiv.org/ftp/arxiv/papers/1601/1601.06971.pdf>. [Accessed: 12- Jun- 2022]
- [6] Alexander Pak and Patrick Paroubek. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA).
- [7] Z. Tong and H. Zhang, "A Text Mining Research Based on LDA Topic Modelling", Computer Science & Information Technology ( CS & IT ), 2016.
- [8] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.
- [9] J. Lee, J. Kang, S. Jun, H. Lim, D. Jang and S. Park, "Ensemble Modeling for Sustainable Technology Transfer", Sustainability, vol. 10, no. 7, p. 2278, 2018.
- [10] A. Hakim, A. Erwin, K. Eng, M. Galinium and W. Muliady, "Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach", 2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE), 2014.
- [11] A. Hakim, A. Erwin, K. Eng, M. Galinium and W. Muliady, "Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach", 2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE), 2014.



[12] Mikheev, Andrei, Marc Moens, and Claire Grover. "Named entity recognition without gazetteers." Ninth Conference of the European Chapter of the Association for Computational Linguistics. 1999.

[13] "GitHub - explosion/spaCy: 🚀 Industrial-strength Natural Language Processing (NLP) in Python", GitHub, 2022. [Online]. Available: <https://github.com/explosion/spaCy>. [Accessed: 12-Jun- 2022]

[14] Spacy.io, 2022. [Online]. Available: <https://spacy.io/models/en>. [Accessed: 12- Jun- 2022]

[15] "GitHub - kevinzg/facebook-scraper: Scrape Facebook public pages without an API key", GitHub, 2022. [Online]. Available: <https://github.com/kevinzg/facebook-scraper>. [Accessed: 12- Jun- 2022]

[16] "GitHub - codelucas/newspaper: News, full-text, and article metadata extraction in Python 3. Advanced docs:", GitHub, 2022. [Online]. Available: <https://github.com/codelucas/newspaper>. [Accessed: 12- Jun- 2022]

[17] N. stopwords, "NLTK's list of english stopwords", Gist, 2022. [Online]. Available: <https://gist.github.com/sebleier/554280>. [Accessed: 12- Jun- 2022]

[18] Tartarus.org, 2022. [Online]. Available: <https://tartarus.org/martin/PorterStemmer/def.txt>. [Accessed: 12- Jun- 2022]

[19] "Welcome to Flask — Flask Documentation (2.1.x)", Flask.palletsprojects.com, 2022. [Online]. Available: <https://flask.palletsprojects.com/en/2.1.x/>. [Accessed: 12- Jun- 2022]

[20]"flask-mail — Flask-Mail 0.9.1 documentation", Pythonhosted.org, 2022. [Online]. Available: <https://pythonhosted.org/Flask-Mail/>. [Accessed: 12- Jun- 2022]