# ARI2201 Individual Assigned Practical Task

## Location Chronicles

Matthew Kenely

matthew.kenely.21@um.edu.mt

**Abstract**

*This project provides a brief overview of the concept of tailoring news articles to users' locations, and the lack of such a platform which tailors Maltese users on a per-village basis. A dataset of Maltese news articles exhaustive of the majority of Maltese villages was created, a Named Entity Recognition model capable of identifying where news articles took place, and a web application which utilises this model to identify the locations in which news articles took place, as well as gives users the ability to filter a database of news articles based on geographical locations was developed using PHP, CSS, JS and Python. The results of this project are discussed and future work in this area is proposed.*

## 1 Introduction

The influx of data in recent years has contributed to the quality of news platforms significantly, with the provision of personalised news content based on users' locations and preferences becoming more accurate by the day. These platforms often focus on larger geographical regions such as countries or cities. For users in smaller regions such as Malta, however, there is a lack of localised news content that caters to their specific interests and preferences on a per-village basis.

Malta has a unique landscape with distinct villages, each with its own local events, issues, and news. For the Maltese population, having access to news articles that are specifically relevant to their local villages can greatly enhance their news consumption experience and keep them informed about happenings in their immediate surroundings.

The aim of this project is to provide Maltese users with a platform which facilitates the recognition of where news articles took place geographically, as well as the ability to filter a database of news articles based on geographical locations, either tailored to a user's location or to whichever village they select. This entails the creation of a news article database, as well as the training of a machine learning model to recognise locations in news articles.

The task of recognising locations in news articles is a Natural Language Processing (NLP) problem, and more specifically, a Named Entity Recognition (NER) problem. NLP is a field of computer science which deals with the interaction between computers and human (natural) languages. NER is a subtask of NLP which deals with the identification of named entities (subjects of interest) in text, such as people, locations and organisations [1]. In this project, the task of the developed NER algorithm will be to identify the locations (specifically, Maltese villages) in which news articles took place.

To successfully create this platform, the following tasks shall be carried out:

1. The creation of a **dataset** of news articles which take place in (and mention) Maltese villages.

2. The training of a **Named Entity Recognition model** on this dataset, with the goal of it being able to recognise where previously unseen Maltese news articles took place.

3. The development of a **web application** on which the following features will be hosted:

    - An article location detector which utilises the trained model to identify locations in news articles provided by the user.

    - A news article database lookup which allows the user to filter said articles in the created dataset based on where they took place in Malta, either tailored to their location (given their consent) or to whichever village they select.

## 2 Background

In recent years, the tendency has been for people to consume news on online platforms [2] [3], offering users a vast amount of news articles from various sources. One popular example of such a platform is Google News, a news aggregator that collects headlines and snippets from news sources worldwide, allowing users to access a wide range of news articles on different topics. One key feature of Google News relevant to this project is its ability to tailor news articles to users' locations. This feature works well when it comes to filtering news articles on a per-country basis (including Malta), and while the option to filter on a per-village basis is available, it is not refined, with Google News often returning any articles which took place in Malta rather than in a specific village.

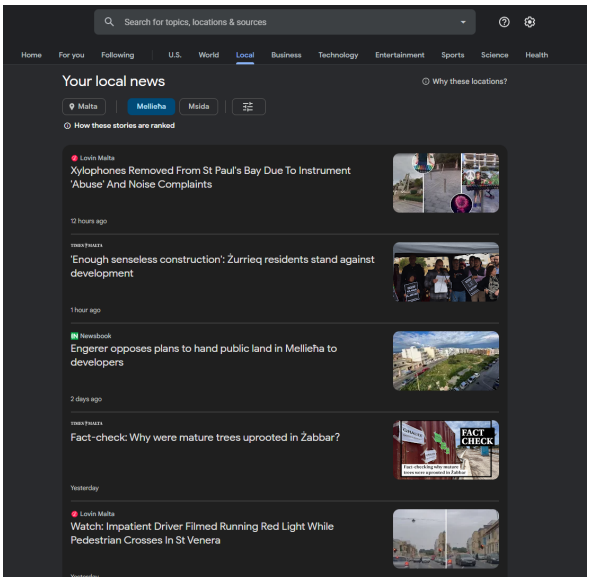


Figure 1: Google News filtering by village (Mellieħa) displaying articles which took place in St. Paul's Bay, Żurrieq and Żabbar

BBC News also offers this feature to a limited extent in its World section, allowing users to filter news articles on a per-continent basis. However, this feature is not available for smaller geographical regions, and as such, cannot tailor news to a user based on the country they live in, let alone the specific city/village.

## 3 Methodology

### 3.1 Dataset

*All source files for work related to this section can be found in the* `docs/data` *directory*

To create a dataset of Maltese news articles, I created algorithms (found in `scraper.ipynb`) which scrape news articles in the Newsbook, MaltaToday and The Malta Independent English local archives. I was able to scrape the following content of 14,177 news articles: URL, title, date, article text. I proceeded to clean this dataset (specifically removing duplicates and normalising date formats) in Google Sheets, followed by a filtering process to only include articles which mention the names of Maltese villages (found in `villages.txt`)

in their text, as well as the start and end index of these villages within the text (this is required to train the NER model), resulting in a dataset of 5226 articles The code for this filtering process can be found in `location_extractor.ipynb`. A final pass through of the dataset was carried out to get the URL of each article's corresponding image (if any) and add it to the dataset (this was done in `scraper.ipynb`).

## 3.2 Model

*All source files for work related to this section can be found in the* `docs/data` *directory*

The Named Entity Recognition model for this project was trained using the spaCy library in Python. spaCy is an open-source library for NLP written in Python, which provides a wide range of NLP tools and features, including the ability to train custom NER models.

I used the blank English model provided by spaCy as a base model (with the intention of creating a model targetted towards the recognition of the names of Maltese villages), and proceeded to train it on the `location_articles_images` dataset, taking the news article text and corresponding village name indices and entity labels (scraped in the Dataset subsection) as inputs. The model was trained for 10 epochs, with a dropout rate of 0.3.

The code used to train the model can be found in `train_model.ipynb`. The model was saved in the `model` directory.

## 3.3 Web Application

*All source files for work related to this section can be found in the* `docs` *directory*

The web application for this project was developed using PHP, CSS, JavaScript and Python. The application must be hosted (this can be done locally using open-source software such as XAMPP).

The primary functions of the application (article location detection and database lookup) are implemented using forms which make POST requests to the server, pass POST parameters to external PHP scripts which run their corresponding Python scripts and return the results of the NER model and database filter respectively.

The application is mobile responsive, allowing users to access it from whichever device they prefer.

The application makes use of the OpenStreetMap API to display a map of Malta on the home page, as well find the coordinates of the villages identified by the Article Location Detector and subsequently display said villages on the map. The Navigator Geolocation API is used on the the News page to ask the user for permission to access their location, which, if provided, is used to automatically filter the news articles in the database based on the village the user is currently in.

### 3.3.1 Home Page

The home page contains a form (called the Article Location Detector) which allows the user to input a news article URL, which is then passed to the NER model to attempt to identify the Maltese village in which the article took place.

### 3.3.2 News Page

The news page contains a form (called the Database Lookup) which allows the user to either provide their current location or select a village manually, which is then passed to a python script which filters the `location_articles_images.csv` dataset, and retrieves news articles which mention that village. I opted for this approach in the case of database retrieval as while news articles may not have necessarily taken place in a user's chosen location, the fact that they mention said location may indicate that the article is of interest to individuals from that location regardless. The articles are then displayed on a grid of cards, each containing the article title, date, image (if any), and a link to the original article. The cards are sorted by date, with the

most recent articles appearing first.

# 4 Results

## 4.1 Dataset

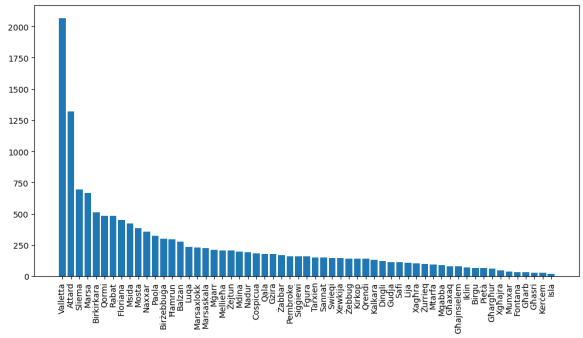

Figure 2: Frequency of villages in the dataset

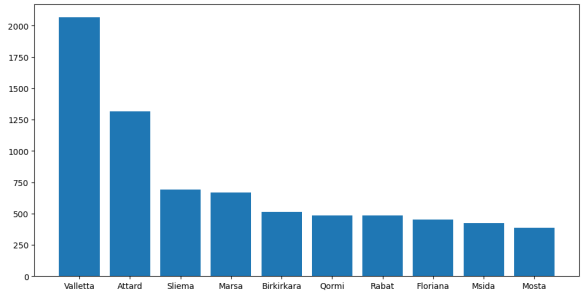


Figure 3: Top 10 most frequently appearing villages

As can be observed in the above figures, the dataset contains news article which occurred in almost all Maltese villages, with most articles taking place in Valletta. This is to be expected given that Valletta is the capital city of Malta and is a very active village.

The dataset is not undersampled in any way and hence is biased towards more active villages. This is not an issue when it comes to the training of an NER model as the linguistic contexts in which the names of villages occur are the same regardless of the popularity of the village, and hence the model will be able to recognise the names of villages (and generalise the contexts in which they appear) regardless of how often they occur in the dataset. Users will simply be shown less articles for less active villages when using the database filter feature.

## 4.2 Model

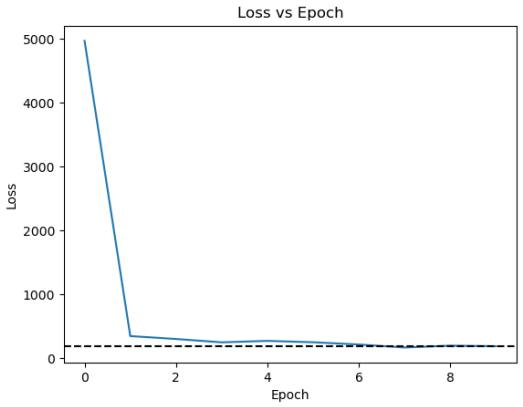


Figure 4: Loss of the NER model over 10 epochs

The following are series of test news articles which were passed to the NER model, where they took place, and the village which the model identified as the location (if any) in which the article took place.

1. Mellieha councillor urges MPs to block development in Natura 2000 site

   → Took place in: **Mellieha**

   → Predicted location: **Mellieha**

2. 'A sea of people' join Puttinu Cares in 17km charity walk

   → Took place in: (Ambiguous)

   → Predicted location: **Mosta**

3. Żurrieq residents' last stand against plans to zone Nigret farmland for apartments

   → Took place in: **Żurrieq**

   → Predicted location: **Żurrieq**

4. Pothole luck: 400 drivers claim compensation for bad roads in two years

   → Took place in: mainly **Siġġiewi**

   → Predicted location: **Siġġiewi**

On the whole, the model performs well, correctly identifying the news artcile location in tests 1, 3 and 4.

In test 2, the news article took place in multiple villages (Mellieha, Xemxija, St Paul's Bay, Mosta, Lija, Birkirkara, Msida, Pieta and Floriana), though it can be argued that primary villages of interest are Mellieha and Floriana. The model is, however, unable to identify either of these as the location in which the article "took place".

### 4.3 Web Application

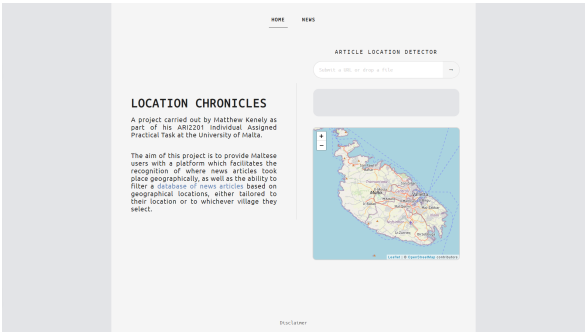


Figure 5: Web application home page

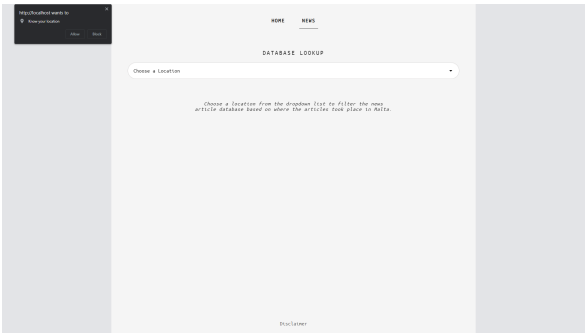


Figure 6: Web application news page (user is prompted for location)

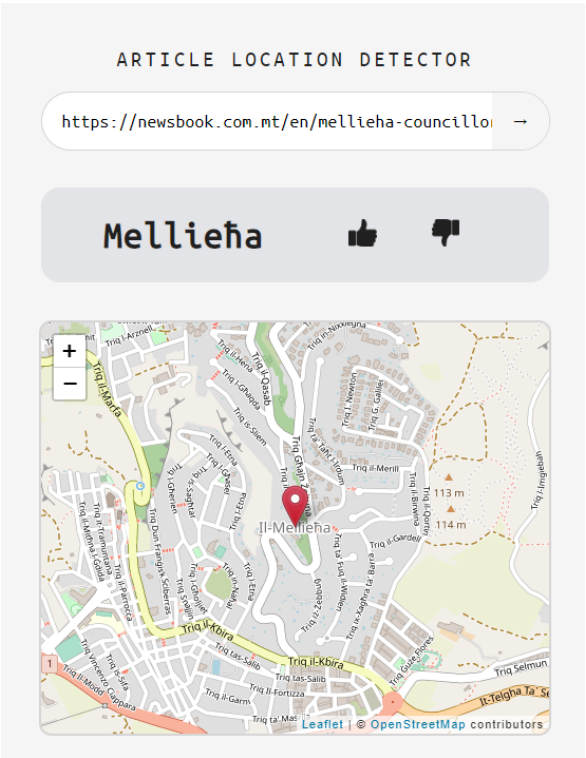


Figure 7: Article location detector

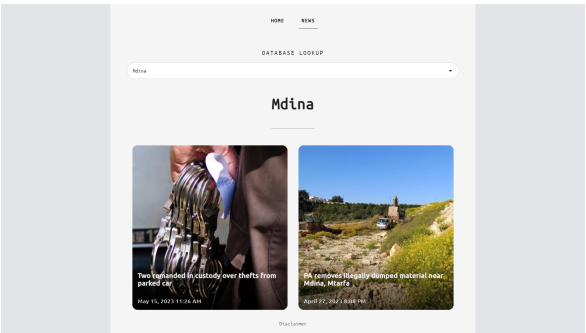


Figure 8: Web application news (user has chosen Mdina manually)

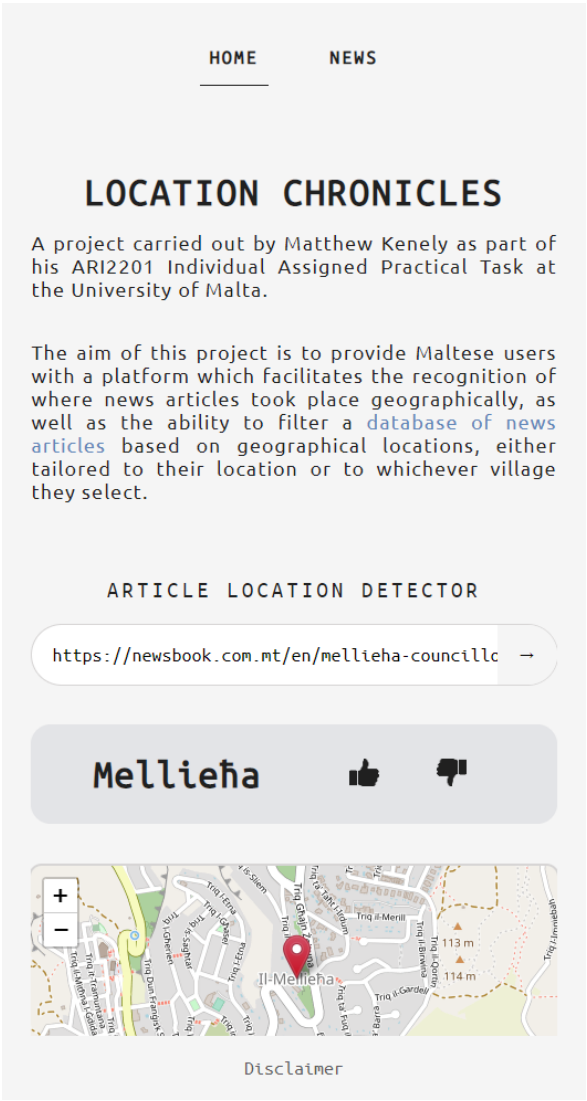


Figure 9: Web application home page on mobile

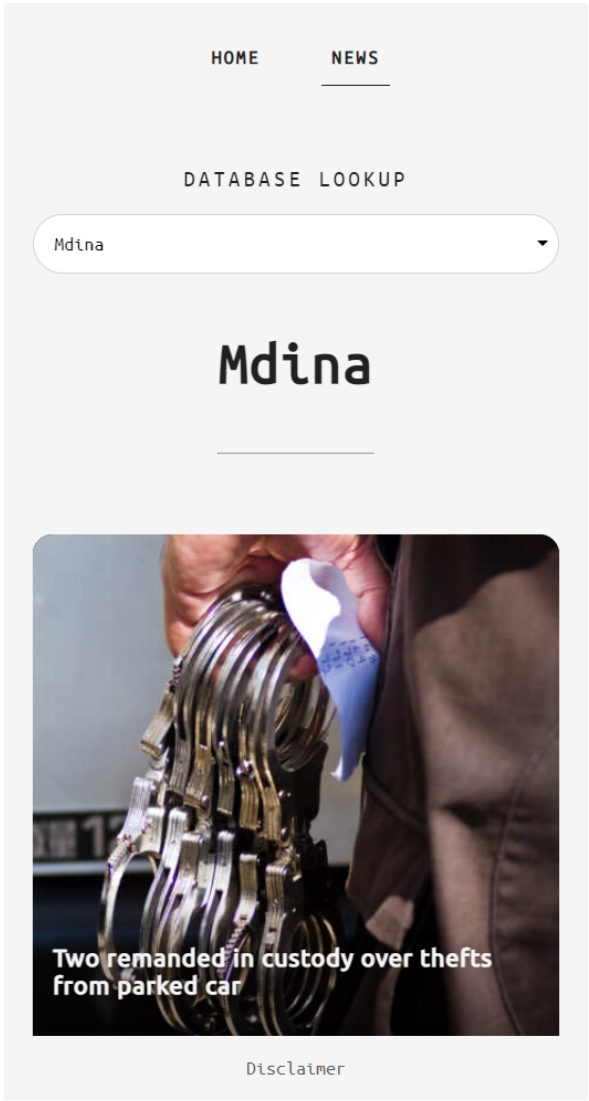


Figure 10: Web application news page on mobile

As can be observed in the above figures, the web application is functioning as intended, with the Article Location Detector and Database Lookup forms successfully making POST requests to the server, calling their corresponding Python scripts and returning the results of the NER model and database filter respectively.

In figure 5, the user is prompted to share their location. If they choose to do so, the application automatically filters the news articles in the database based on the village they are currently in. If they choose not to share their location, they are prompted to select a village manually, as shown in figure 7.

The application is also fully mobile re-

sponsive, allowing users to access it from whichever device they prefer.

## 5 Challenges and Limitations

The dataset only contains English news articles. This is due to the fact that the NER model was trained on the English spaCy model, and hence is only able to recognise locations in English text. This is a limitation of the model, and can be addressed by training the model on a dataset of Maltese news articles, and using a Maltese NLP library such as

The model is unable to recognise village names which contain multiple words, such as St. Paul's Bay. This is due to the fact that the model was trained on a dataset of news articles which only contain single-word village names. This is a limitation of the model, and can be addressed by training the model on a dataset of news articles which contain multi-word village names, and labelling the village names in the text as such.

The model is unable to recognise the names of villages in news articles which took place in multiple villages. This is a limitation of the model, and can be addressed by training the model on a dataset which contains news articles which took place in multiple villages, and labelling the villages in which they took place as such.

## 6 Conclusion

In this project, I created a dataset of Maltese news articles, trained a Named Entity Recognition model on this dataset, and developed a web application which utilises this model to identify the locations in which news articles took place, as well as filter a database of news articles based on geographical locations.

Future work on this project could include the following:

- Training the NER model on a dataset of Maltese news articles, and using a Maltese NLP library such
- Training the NER model on a dataset of news articles which contain multi-word village names, and labelling the village names in the text as such.
- Training the NER model on a dataset which contains news articles which took place in multiple villages, and labelling the villages in which they took place as such.

## References


[1] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[2] S. Bennett, K. Maton, and L. Kervin, "The 'digital natives' debate: A critical review of the evidence," *British journal of educational technology*, vol. 39, no. 5, pp. 775–786, 2008.

[3] A. C. Ripollés, "Beyond newspapers: News consumption among young people in the digital era," *Comunicar. Media Education Research Journal*, vol. 20, no. 2, 2012.