# **Intro**

In a news story, recent or current events are covered, either generally (as in daily newspapers) or in relation to a particular subject (i.e. political or trade news magazines, club newsletters, or technology news websites). Reports from people who witnessed an occurrence can be included in news articles. When we visit a news website, we must have seen that the news is separated into categories. technology, entertainment, sports, and other popular categories can be found on practically all news websites.

Before releasing an item, every news website organises it into categories so that users may quickly select the categories of news that interest them.

For instance, I frequently visit news websites, blogs and click on the technology section because I want to read about the most recent technological developments. You might prefer to read about politics, business, entertainment, or even sports if you don't enjoy reading about technology. The content administrators of news websites currently classify the news stories by hand. However, in order to save time

they can also incorporate a machine learning model on their websites that reads the news headline or the news's content and categorises it and summarise the news. So this falls into our project core.

**Life Cycle of Machine Learning Project**

Life Cycle of implementing machine learning application.

* Gathering the Data
* Data Preparation and Feature Selection
* Feature Selection
* Create Model
* Evaluate Model
* Deploy the model

1. **Gathering the Data**
   1. **Using Web-scraping framework from news**

[BOOM FactCheck: Debunking Misinformation (boomlive.in)](https://www.boomlive.in/)

[Snopes.com - FactCheck.org](https://www.factcheck.org/2009/04/snopescom/?gclid=Cj0KCQjwxveXBhDDARIsAI0Q0x1zxOlaG1JtoBYSvarW9TInplN4p5MHudXu_ntMSQ-crhgQXgz_OVYaAqbiEALw_wcB)

[PolitiFact](https://www.politifact.com/)

[AllSides | Balanced news via media bias ratings for an unbiased news perspective](https://www.allsides.com/unbiased-balanced-news)

* 1. **From the article data-set**

[BBC News Classification | Kaggle](https://www.kaggle.com/competitions/learn-ai-bbc/data)

* 1. **Direct Text**
  2. **From Image**

Tesseract, pytesseract, pillow

1. **Data Preparation and feature extraction**

* Count vectorization, Bag of Words
* Term frequency, inverse document frequency
* Stop word removal
* Feature Selection

1. **Feature Selection**

* Chi-squared
* Least Absolute Shrinkage and Selection Operator (LASSO)

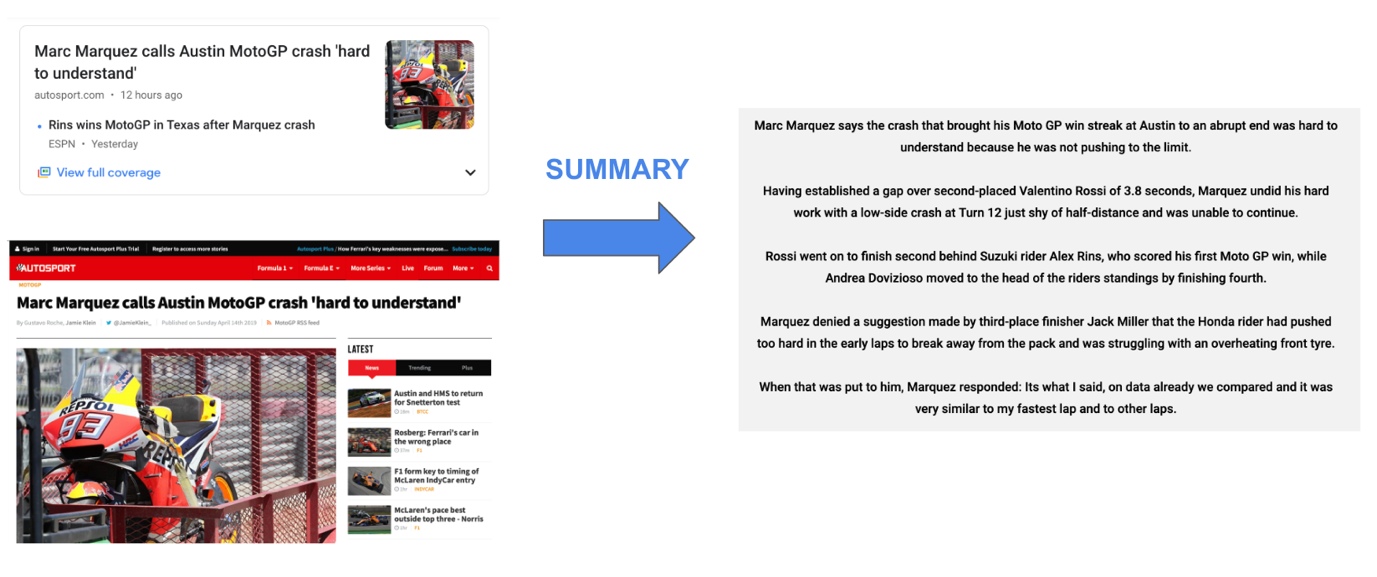
# **Text Summarising**

What is Text summarising ???

Text summarising is a strategy for producing a brief, accurate summary of lengthy texts while concentrating on the passages that provide relevant information and keeping the overall meaning intact.

The goal of automatic text summarising is to reduce long papers to their essential content. Manual text summarization could be time-consuming and expensive.

Before creating the necessary summary sentences, machine learning algorithms can be trained to understand documents and recognise the sections that carry key facts and information.



Need for Text summarising ???

The current expansion of non-structured textual data in the digital sphere necessitates the creation of automatic text summary technologies that make it simple for users to draw conclusions from them. We now have immediate access to vast volumes of information. However, the majority of this data is unnecessary, trivial, and might not convey the intended meaning.   
For instance, if you are searching for certain information in a news item online, you may need to sift through the post's content and spend a lot of time eliminating the extraneous information before finding what you are looking for. Therefore, it is increasingly important to use computerised text summarizers that can extract useful information while excluding inessential and irrelevant data. Utilizing summary can improve readability.

**Extraction based summarization**  
In extraction-based summarization, a subset of words that represent the most important points is pulled from a piece of text and combined to make a summary. Think of it as a highlighter—which selects the main information from a source text.

In machine learning, extractive summarization usually involves weighing the essential sections of sentences and using the results to generate summaries.

Different types of algorithms and methods can be used to gauge the weights of the sentences and then rank them according to their relevance and similarity with one another—and further joining them to generate a summary.

Here's an example:  
Peter and Elizabeth took a taxi to attend the night party in the city.  
While in the party, Elizabeth collapsed and was rushed to the hospital.

Summary : Peter and Elizabeth attended party city. Elizabeth rushed hospital.

**Abstraction based summarization**In abstraction-based summarization, advanced deep learning techniques are applied to paraphrase and shorten the original document, just like humans do. Think of it as a pen which produces novel sentences that may not be part of the source document.

Since abstractive machine learning algorithms can generate new phrases and sentences that  
represent the most important information from the source text, they can assist in  
overcoming the grammatical inaccuracies of the extraction techniques.

Here's an example:  
Peter and Elizabeth took a taxi to attend the night party in the city.  
While in the party, Elizabeth collapsed and was rushed to the hospital.

Summary 🡪 Elizabeth was hospitalized after attending a party with Peter.

## Implementation

**Implementation for extractive method**

Gensim with TextRank, Text Summarization with Sumy**,** LSA (Latent semantic analysis)**,** Luhn **,** KL-Sum

**Implementation for Abstractive method**

Huggingface’s transformers, T5 Transformers, BART Transformers, GPT-2 Transformers, Open AI Transformers

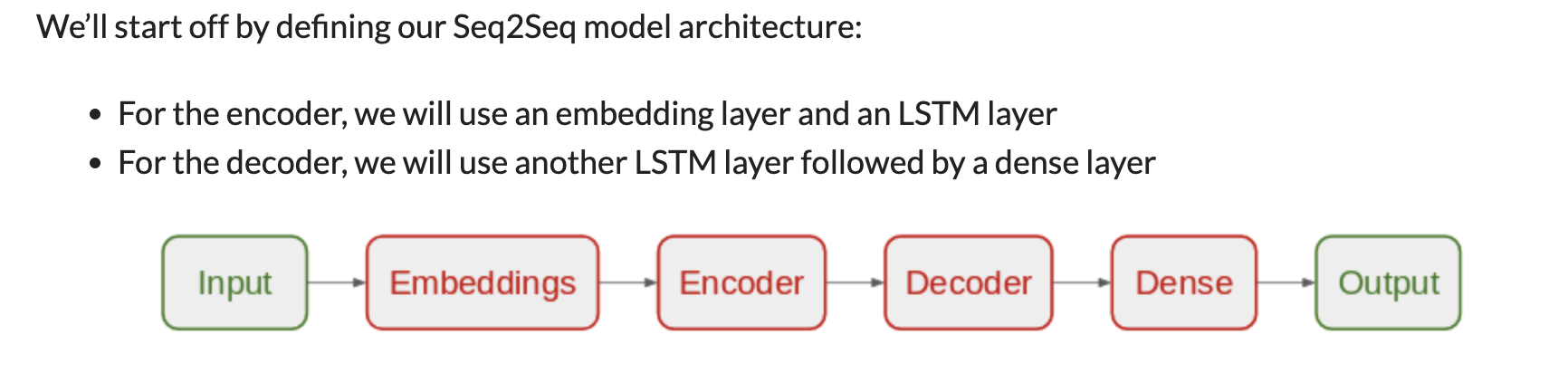
# **Text Translator**

* Optical Character Recognition (OCR) is a computer-based electronic method for converting text from image files into machine-encoded text that can be extracted and used in text format.
* An open source OCR (optical character recognition) engine called Tesseract enables text to be extracted from photos.
* A wrapper for the Tesseract engine is the Pytesseract library.

## Implementation

🡪 Specifying Source / Identify the language source and Destination Language

🡪 Using translator object, translate the text



# **News Categorization**

When you visit a news website, you may have noticed that the news is segmented into categories. Tech, entertainment, and sports are a few of the popular categories that you'll find on practically any news website.

Before releasing an item, every news website organises it into categories so that users may quickly select the categories of news that interest them. For instance, I frequently visit news websites and click on the technology section because I want to read about the most recent technological developments. You might prefer to read about politics, business, entertainment, or even sports if you don't enjoy reading about technology.

The content administrators of news websites currently classify the news stories by hand. However, in order to save time, they can also incorporate a machine learning model on their websites that reads the news headline or the news's content and categorises it.

## Implementation

**Classifier Methods**a) Naïve Bayes  
b) Support Vector Machine (SVM)  
c) Artificial and Recurrent convolutional neural networks

# **References**

<https://stackabuse.com/text-translation-with-google-translate-api-in-python/>

<https://www.analyticsvidhya.com/blog/2021/12/text-classification-of-news-articles/>

<https://www.analyticsvidhya.com/blog/2021/12/text-classification-of-news-articles/>

<https://iopscience.iop.org/article/10.1088/1757-899X/459/1/012006/pdf>

Language translation

<https://towardsdatascience.com/how-to-detect-and-translate-languages-for-nlp-project-dfd52af0c3b5>

<https://www.analyticsvidhya.com/blog/2019/01/neural-machine-translation-keras/>

Text Extraction from web-site

<https://towardsdatascience.com/scraping-1000s-of-news-articles-using-10-simple-steps-d57636a49755>

Image to Text

<https://towardsdatascience.com/extract-text-from-image-using-python-8e8cfbbce743>