



RUTGERS BUSINESS SCHOOL,  
RUTGERS UNIVERSITY,  
NEWARK

MASTER OF INFORMATION TECHNOLOGY AND  
ANALYTICS  
CAPSTONE PROJECT

**TOPIC: ONE STOP SOLUTION FOR NEWS ARTICLES**  
**(CLASSIFICATION AND RELIABILITY CHECK (FAKE**  
**NEWS DETECTION))**

A project submitted in partial fulfillment  
of the requirements of the  
Master of Information Technology and Analytics coursework.

Darsh Thakkar  
December 2020

PRESENTED BY: DARSH THAKKAR  
RUID: 196004707

ADVISOR: PROFESSOR SERGEI SCHREIDER

## **Acknowledgements**

I would like to take the time and opportunity to thank my advisor for this Capstone Project- Professor Sergei Schreider. Throughout my master's degree, I have had the opportunity to take 3 courses under Professor Schreider's supervision- namely Analytics for Business Intelligence, Business Forecasting and Capstone Project. He has guided me in the best way possible, throughout my journey at Rutgers. This project wouldn't have been possible without his help and guidance.

I would also like to thank the teaching and non-teaching staff of Rutgers University, who have been a constant source of help throughout this year and a half, especially during the time of this pandemic. Rutgers University has taken all the necessary steps to maintain the quality of education provided and has taken extra steps to make students' lives easier.

Lastly, I'd like to thank my family and friends for believing in me and for the wonderful journey that I have had as a master's candidate.

## Contents

<b>1. Introduction.....</b>	<b>5</b>
<b>2. Supervised Learning Approach.....</b>	<b>6</b>
<b>3. News Classification Model.....</b>	<b>7</b>
a) Tools.....	7
b) Model Architecture.....	8
c) Data Dictionary.....	9
d) Data Cleaning, Loading and Feature Extraction.....	10
e) Model Performance Evaluation.....	12
f) Model Deployment Cost.....	13
<b>4. Fake News Model Detection.....</b>	<b>14</b>
a) Tools.....	15
b) Model Architecture.....	16
c) Data Dictionary.....	17
d) Data Cleaning, Loading and Feature Extraction.....	18
e) Model Performance Evaluation.....	20
<b>5. Conclusion.....</b>	<b>21</b>
<b>6. Future Scope.....</b>	<b>22</b>
<b>7. References.....</b>	<b>23</b>

## Introduction

‘Data’ has taken world by storm. Humans generate massive amounts of data everyday- knowingly and unknowingly. This is what the Scientists term as ‘The Data Glut’. Pulling out some facts and figures:

- **1.7MB of data** is created every second by every person during 2020.
- In the last two years alone, we managed to create **90%** of the data that exists today.
- **2.5 quintillion bytes** of data are produced by humans every day.
- **95 million** photos and videos are shared every day on Instagram.
- Every day, **306.4 billion emails** are sent, and **5 million Tweets** are made.
- It is predicted that **44 zettabytes** of data will make up the entire digital universe, by 2020.
- And a mammoth **463 exabytes** of data will be generated each day by humans as of 2025.

What is happening due to this is that tremendous amount of important data and information is either being lost or being fabricated. The fabrication of information leads to mass-panic, superficial believes and biased judgements. And this information is mostly fabricated in the form of news. This is what we call fake news.

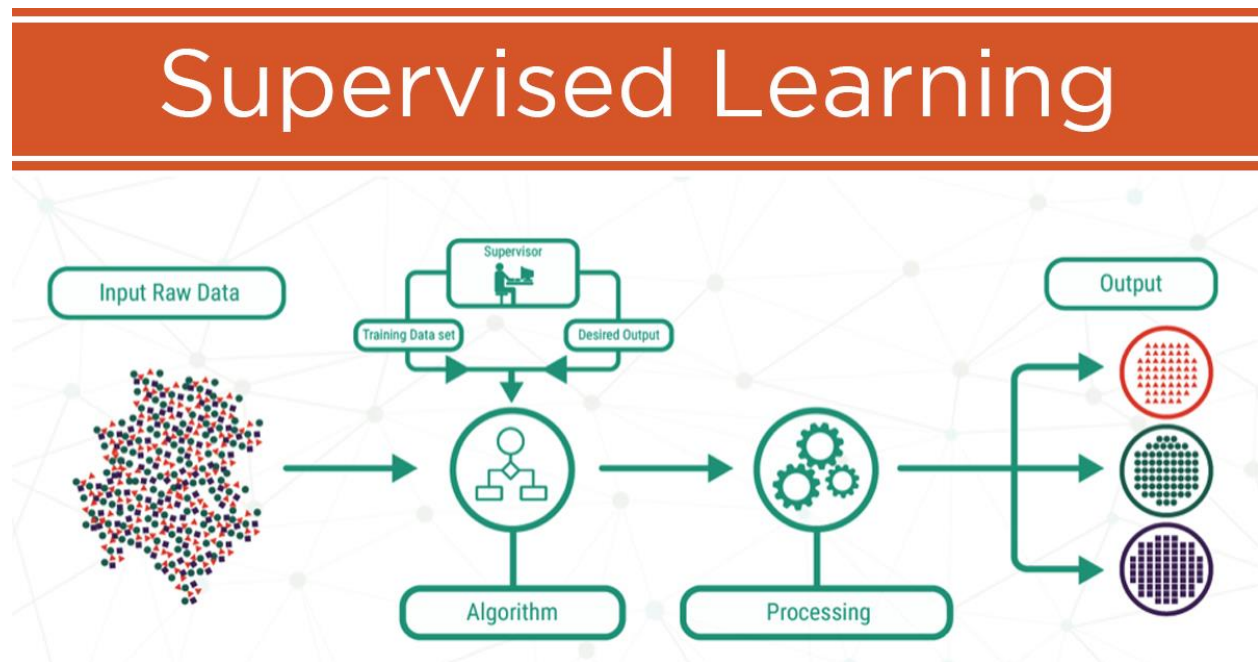
The motivation behind developing and completing this project was to efficiently deliver a one stop solution to classify news and also to efficiently tackle fake news using the best tools and technologies in Natural Language Processing and Classification.

## Supervised Learning Approach

Both, The News Classification as well as The Fake News Detection model use Supervised Learning Approach.

Supervised learning is a machine learning domain that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples.

The architecture of Supervised Learning is as given below:

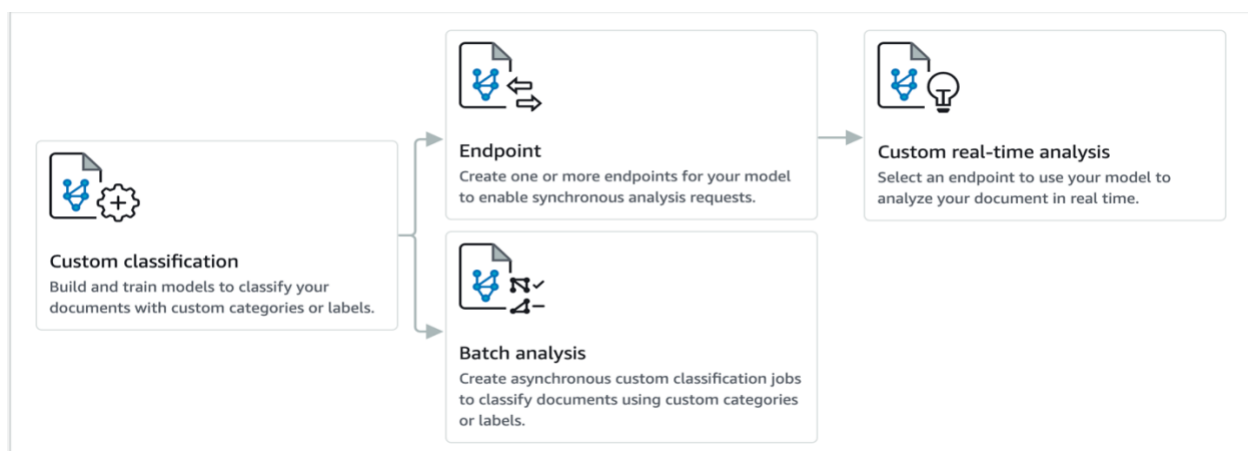


Both the models use labelled data to train the machine to make accurate predictions/classifications when new instances of data are passed to the model.

## News Classification Model Tools

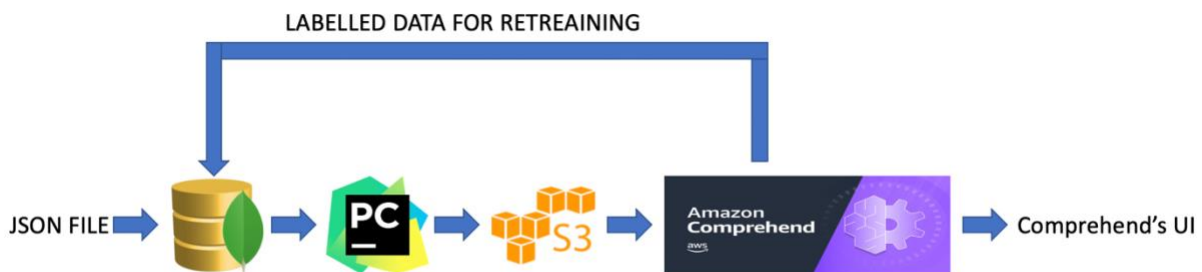
The following tools and technologies were used to create an end-to-end pipeline and build a production level for the model for this use case:

- **Terminal:** Terminal is used to write Linux Shell Commands in a Macintosh Computer. Ruby and Homebrew commands are essential to operate a MongoDB database.
- **MongoDB:** MongoDB is a database, which has both cloud and local storage options. The objects in MongoDB are stored in a BSON (Binary JavaScript Object Notation) format.
- **PyCharm:** PyCharm is an Integrated Development Environment, which here is used with Python 3.8 scripting language to pull the data, pre-process it and feed it into a CSV file.
- **AWS S3 Buckets:** The AWS S3 bucket stores the CSV file, which is necessary to train using Amazon Comprehend.
- **Amazon Comprehend:** Amazon Comprehend is a Machine Learning tool that can perform various custom classification tasks like detecting dominant language, key entity recognition etc. It also provides real-time and batch-analysis options to analyze the input data and generate results with confidence intervals. The image below shows a high-level overview of Amazon Comprehend's custom classification architecture.



## Model Architecture

The pipeline architecture for model is as shown below:



The Dataset for this model is derived from Kaggle, and can be found on the following link: <https://www.kaggle.com/rmisra/news-category-dataset>

This dataset was downloaded in JSON format, which was then converted into a pandas data-frame.

It has columns for category, headline, author, link, short description and date.



## Data Dictionary for News Classification Model

Name of The Columns	Description
<b>category</b>	The category which the news article falls into. For example: Crime, Sports, etc.
<b>headline</b>	The headline of the news article.
<b>authors</b>	The author who wrote the news article.
<b>link</b>	The link for the news article.
<b>link</b>	The link for the news article.
<b>date</b>	The date when the news article was written.

## Data Cleaning, Loading and Feature Extraction

The problem with data was that it was scrapped from multiple websites and each website had its own terminology for the categories.

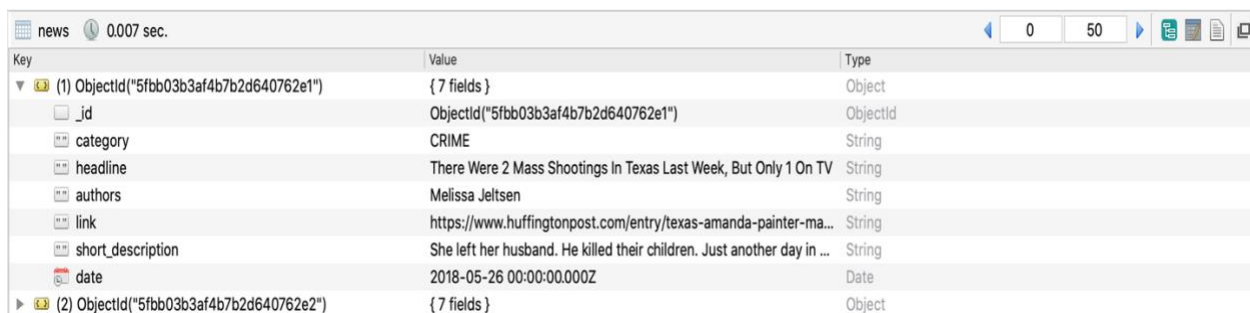
Research was carried out for best practices, by visiting the news websites like- The WSJ, New York Times and it was decided to narrow the data down to 10 categories. The categorization for WSJ is as shown below for reference.

# THE WALL STREET JOURNAL.

English Edition | Print Edition | Video | Podcasts | Latest Headlines

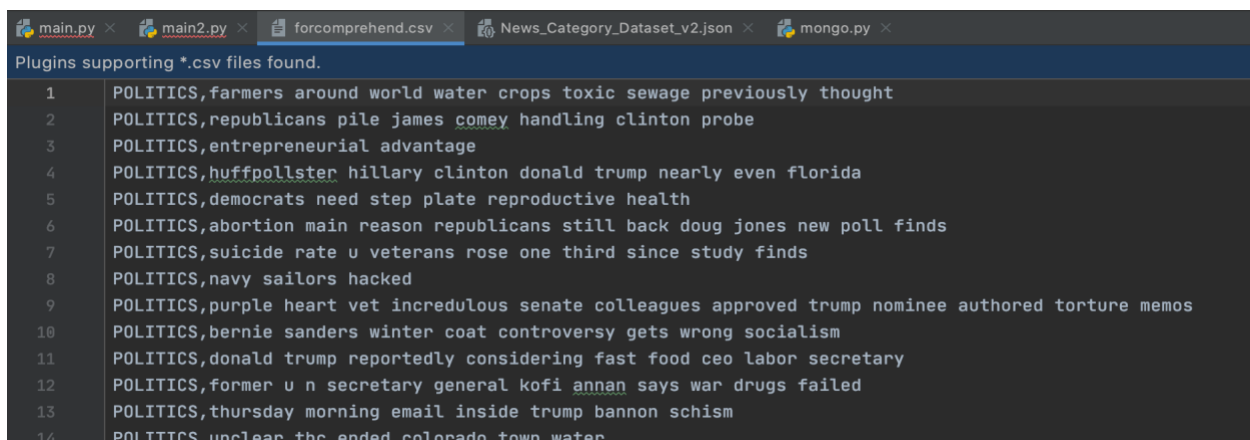
Home World U.S. Politics Economy Business Tech Markets Opinion Life & Arts Real Estate WSJ Magazine

The data-frame is then converted to dictionary object and then pushed into MongoDB Database. MongoDB Database stores the data into BSON format, that is Binary JavaScript Object Notation. And each object has the below format:



Key	Value	Type
(1) ObjectId("5fbb03b3af4b7b2d640762e1")	{ 7 fields }	Object
_id	ObjectId("5fbb03b3af4b7b2d640762e1")	ObjectId
category	CRIME	String
headline	There Were 2 Mass Shootings In Texas Last Week, But Only 1 On TV	String
authors	Melissa Jeltsen	String
link	https://www.huffingtonpost.com/entry/texas-amanda-painter-ma...	String
short_description	She left her husband. He killed their children. Just another day in ...	String
date	2018-05-26 00:00:00.000Z	Date
(2) ObjectId("5fbb03b3af4b7b2d640762e2")	{ 7 fields }	Object

This data is then pulled using python, cleaned and converted into a CSV file.



Plugins supporting *.csv files found.	
1	POLITICS,farmers around world water crops toxic sewage previously thought
2	POLITICS,repblicans pile james comey handling clinton probe
3	POLITICS,entrepreneurial advantage
4	POLITICS,huffpollster hillary clinton donald trump nearly even florida
5	POLITICS,democrats need step plate reproductive health
6	POLITICS,abortion main reason republicans still back doug jones new poll finds
7	POLITICS,suicide rate u veterans rose one third since study finds
8	POLITICS,navy sailors hacked
9	POLITICS,purple heart vet incredulous senate colleagues approved trump nominee authored torture memos
10	POLITICS,bernie sanders winter coat controversy gets wrong socialism
11	POLITICS,donald trump reportedly considering fast food ceo labor secretary
12	POLITICS,former u n secretary general kofi annan says war drugs failed
13	POLITICS,thursday morning email inside trump bannon schism
14	POLITICS,unclear thc ended colorado town water

The CSV file is then pushed into an Amazon S3 bucket. The CSV from there, is then pulled by Amazon Comprehend to perform custom model training.

**Objects (5)**  
Objects are the fundamental entities stored in Amazon S3. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

[Refresh](#) [Delete](#) [Actions](#) [Create folder](#) [Upload](#)

Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	<a href="#">.write_access_check_file.temp</a>	temp	November 26, 2020, 11:45 (UTC-05:00)	0 B	Standard
<input type="checkbox"/>	<a href="#">735053416321-CLR-2f08372b1ebddb24a0fc5012e83efc6/</a>	Folder	-	-	-
<input type="checkbox"/>	<a href="#">735053416321-CLR-603ad3c1e98027a72021e90458a8a39a/</a>	Folder	-	-	-
<input type="checkbox"/>	<a href="#">forcomprehend.csv</a>	csv	November 26, 2020, 11:44 (UTC-05:00)	1.3 MB	Standard
<input type="checkbox"/>	<a href="#">forcomprehend2.csv</a>	csv	November 26, 2020, 10:52 (UTC-05:00)	337.2 KB	Standard

## Model Performance Evaluation

The accuracy achieved during training using Amazon Comprehend was about 92% with a hamming loss of 7.8%.

Classifier performance <a href="#">Info</a>	
Accuracy	0.922
Hamming loss	0.078

The Amazon Comprehend news classification model correctly and accurately classifies the news articles into 10 categories (labels) and also shows the confidence interval for the classification result.

- World
- Tech
- Sports
- Politics
- Parenting
- Opinion
- Health
- Entertainment
- Crime
- Business

The examples for Real-Time Analysis that was performed are shown below:

**Custom**

View real-time insights based on custom models from an endpoint you've created.

**Endpoint**

endend

Custom classifier: finalcapclass

**Input text**

In-Depth: Lakewood, Cleveland police work to solve carjacking crime spree

73 of 5000 characters used.

Clear text

Analyze

**Insights** [Info](#)

**Analyzed text**

In-Depth: Lakewood, Cleveland police work to solve carjacking crime spree

▼ **Results**

**Classes**

CRIME  
0.96 confidence

OPINION  
0.01 confidence

POLITICS  
0.00 confidence

► Application integration

**Insights** [Info](#)

**Analyzed text**

Joe Burrow's injury was 'tough to watch,' LSU coach Ed Orgeron says

▼ **Results**

**Classes**

SPORTS  
0.98 confidence

CRIME  
0.00 confidence

OPINION  
0.00 confidence

► Application integration

## **Model Deployment Cost**

- For asynchronous classification and entity recognition: \$0.0005 PER UNIT (Inference requests are measured in units of 100 characters, with a 3 unit (300 character) minimum charge per request.)
- For synchronous classification and entity recognition: \$0.0005 PER IU PER SECOND (Endpoints are billed on one second increments, with a minimum of 60 seconds. Charges will continue to incur from the time you start the endpoint until it is deleted even if no documents are analyzed.)
- One inference unit (IU) provides a throughput of 100 characters/second on your managed endpoint. You can provision additional IUs for more throughput. Each IU will incur \$0.0005 per second.
- \$3 PER HOUR FOR MODEL TRAINING
- \$0.50 PER MONTH FOR MODEL MANAGEMENT

**The cost of this particular model training came down to \$7. Batch analysis/real-time analysis is billed separately.**