DATA8001 - Assignment 2

### Student ID: R00206995

### Student Name: Neto Ribeiro

# Report

## Summary

I have started using a sample of 100 files to read the news article files, I spent time modeling a structure to read and save all files at the same time and I was happy with the result.

Before creating the multi-class classification models I did research on the pre-processing techniques and I choose the TF-IDF technique and I have applied lowercase, lemmatize and remove the stop words also I have removed special characters and punctuation on the text preparation.

Finally, I did an exploratory data analysis and I found a pattern on the data and I choose to proceed with only NOUNS and ADJECTIVES on top of the prepared data and choose the words with were statistically relevant.

From the histogram plot, we can see the dataset is proportionally balanced of 500 observations for each News Category.

Chart

Description automatically generated

Now we can see the TOP 10 relevant words for News Category:

Text

Description automatically generated

For the first model, I have created a model using SKLEARN SVM model which had an overall accuracy of 91.5%.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Detail:** | **precision** | **recall** | **f1-score** | **support** |
|  |  |  |  |  |
| **ENTERTAINMENT** | 86% | 85% | 86% | 109 |
| **SPORTS** | 88% | 79% | 83% | 99 |
| **TECHNOLOGY** | 74% | 77% | 76% | 94 |
| **WORLD** | 77% | 82% | 79% | 99 |

For the second model, I have created the model using the Naive Bayes Theorem algorithm which had an overall accuracy of 98%.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Detail:** | **precision** | **recall** | **f1-score** | **support** |
|  |  |  |  |  |
| **ENTERTAINMENT** | 92% | 91% | 92% | 108 |
| **SPORTS** | 94% | 85% | 89% | 99 |
| **TECHNOLOGY** | 79% | 83% | 81% | 94 |
| **WORLD** | 82% | 88% | 85% | 99 |

For the third model, I have created the model using the Random Forest Classifier (RFC) which had an overall accuracy of 92%.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Detail:** | **precision** | **recall** | **f1-score** | **support** |
|  |  |  |  |  |
| **ENTERTAINMENT** | 82% | 81% | 82% | 108 |
| **SPORTS** | 83% | 75% | 79% | 99 |
| **TECHNOLOGY** | 70% | 74% | 72% | 94 |
| **WORLD** | 73% | 78% | 75% | 99 |

## Conclusion

I would recommend the Model 2 – Naive Bayes Theorem algorithm for classification model which TF-IDF pre-processed data combined with text preparation method using lowercase, lemmatize and stop words, on top of it filtering only NOUNS and ADJECTIVES and finally get only the most statistically relevant words to apply the [Chi-Squared Test](https://en.wikipedia.org/wiki/Chi-squared_test) technique.

|  |  |
| --- | --- |
| Chart, bar chart  Description automatically generated | Chart, line chart  Description automatically generated |

Text, whiteboard

Description automatically generated

Predicted News Category using the example news headline ("Covid: Brazil passes 400,000 deaths amid slow vaccination")

Using lime text explainer we can get this very good visualization of the predict outcome:

Text

Description automatically generated

# Question 1 - Why do you think that AI projects have such a poor rate of success?

#### (max 300 words)

I would say that AI projects have such a poor rate of success because of two things,

Firstly it is really normal rate when we are talking about IT projects, my background is in ERP for Retailers, and using my own experience I can tell after 15 years working direct with CEO and Directors you can only get approve on your project when you are able to show how much money your project want to save or how much human errors your project will be able to avoid.

Secondly I think the poor rate of success is because of the full understanding around the what AI could do for your company, most of people doesn’t know that answer and sometimes they think the AI would do everything you can image without any commitment from you or from your team, AI needs lots of commitment specially from the one which knows everything about the business and in the most of the time that person are very busy and they don’t have to spend on the new project.

Finally I think to have a success AI project the customer needs to see importance the same as you can see when you are planning to create an AI project from the scratch, you should show that AI is already there on the emails when some emails goes direct to the “SPAM” folder or when you are tipping and your email is giving you next words as option, when your Netflix is suggesting you next film to watch and so long of real life AI project already there.

# Question 2 - Discuss the difference between Schema on Read vs. Schema on write architectures with examples.

#### (max 300 words)

Schema on write is very popular structured table bases, with there was a lots of time spent behind an simple and small project, to create an database you should to starting with design it before, when we are talking about old project or ERP your cannot just decide to add new column on your simple Customer Table because it might causes errors in your program.

Scheme on read have opposite of the scheme on writhe, because you are adding data as you are scrapping or receiving from lots of many sources, however, you need to know exactly what you are looking for otherwise you can easily get lost on your silo.

Most popular scheme on write database is from Microsoft called SQL Server, the data is stored in a relational tables schema, with columns and rows, the columns denotes the attribute and rows are the record, there is a special duty to handle the Database called DBA (Database Administrator), you can fully controlling the data before save in the database, most common example if to not allow NULL or EMPTY values.

Most popular scheme on read is MongoDB and data is stored in collections, A collection consist of many documents in which data is stored in JSON file format of key-value.

You can easily start a project using scheme on read to get insights from your data and also from the new data as an experimental project.