Economic News Identification Using an LSTM Neural Network Approach

Jan Maciejowski, Fabian Perez, Ali Rammal, Louis Golding, Abdullah Ghosheh, Ayah El Barq

Introduction

In an era where information is abundant, researchers, policy makers and executives face an extreme task of distinguishing economically relevant articles. This goal is the primary objective of our project with the use of the dataset (US-Economic_news.csv), our goal is to create a classification model that can identify documents based on how relevant they are to modern economy.

The dataset we have contains several attributes such as positivity, relevance, date and headline and displays various articles. These attributes are pivotal for the project and for the model, it helps us categorize the news based on relevancy.

To start with the project, we have to conduct an in-depth analysis on the attributes to check their efficiency towards the articles and to assess whether they're going to be important for building the model or not, by doing so we can receive insights on the attributes efficiency. Secondly, with such insights we continue to build the classification model that can assess articles and determining whether they're relevant or not.

In [18]:

```
import pandas as pd
from ydata_profiling import ProfileReport
import matplotlib.pyplot as plt
```

1: EDA and Preprocessing

Initial EDA

This notebook contains the code to load the "US economy news" dataset

In [2]:

```
df_news = pd.read_csv('US-Economic-News.csv', delimiter=',', encoding = 'ISO-885
print(df_news.columns)
print()
print(df_news.shape)
```

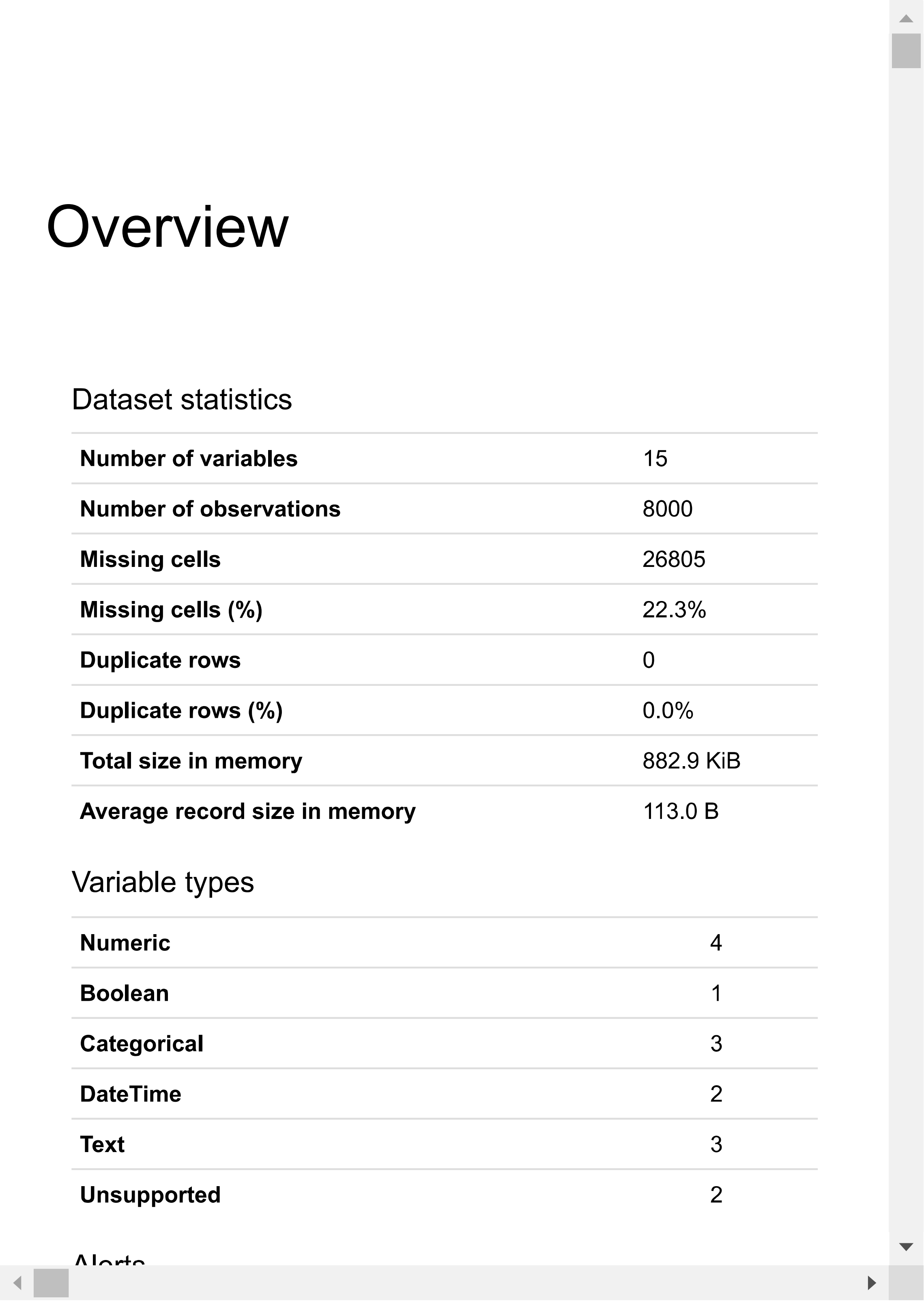
```
Index(['_unit_id', '_golden', '_unit_state', '_trusted_judgments',
       '_last_judgment_at', 'positivity', 'positivity:confidence', 'relevance',
       'relevance:confidence', 'articleid', 'date', 'headline',
       'positivity_gold', 'relevance_gold', 'text'],
      dtype='object')

(8000, 15)
```

```
In [ ]: df_news.head(20)
```

```
In [4]: profile = ProfileReport(df_news, title="News Profile Report")
        profile.to_file('your_report.html')
        profile
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]
Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]
Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]
Export report to file:   0%|          | 0/1 [00:00<?, ?it/s]
```



Overview

Dataset statistics

Number of variables	15
Number of observations	8000
Missing cells	26805
Missing cells (%)	22.3%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	882.9 KiB
Average record size in memory	113.0 B

Variable types

Numeric	4
Boolean	1
Categorical	3
DateTime	2
Text	3
Unsupported	2

Alerts

Out[4]:

From our data profiling, we can see that we have missing values in our dataset. 22.3% of our data is missing. Exploring our dataset will help us understand the nature of the variables available and help us decide what to do in terms of preprocessing

```
target variable:
    relevance: relevance is the variable that is either yes or
no, which is either economically relevant or irrelevant

id:
    The unit_id refers to the file id of obtained features of
determining the relevance of an article.

constants:
    Golden is a constant variable so we should not include it in
our data for the model.
    _unit_state is also a constant variable so it would not
provide useful information in our models.
    _trusted_judgments is another constant variable that does
not provide variance

    We would assume that relevance gold and positivity gold are
connected to golden. They are missing and corrupt, and removing
golden because of this other reason seems correct

variables:
    positivity: highly correlated with the target variable,
however there are 82.2% missing values. In the profile, the
histogram shows that curve is not normal, so imputing with the
mean could be a good strategy. At the same time, we do not know
what this variable means. We thought it could mean the positive
words in the article, or whether the article was liked as higher
scores = relevant. Leaving this variable out of the features for
the model would be a better idea as we would save more time and
be efficient. A neural network is able to consume text and make
classifications solely based on that, through the use of
different text processing libraries and methodologies

    positivity confidence: there was no information to conclude
the value of this data. it is also missing, so it would be best
to keep it out of the features.

    relevance confidence: there is no correlation and the
context of this data is not understood. We will not include it
in our features

    date: we will not use the date of publication as there is no
correlation to relevance-no patterns identified.

    article id: article identifier
```

From this, we will only be using the text and headlines as features for our model building. We will explore different models with an endgoal of building a neural network that is able to classify whether a new article is economically relevant or not. In our preprocessing, we can try unique ways of tokenizing and vectorizing the text, decide to keep headlines only or include the whole text. There are a number of methodologies that can be pursued to produce a capable neural network with high accuracy on test set.