Week 7 Report

**Name: Amr Elbana**

**Email: amr32009363@gmail.com**

**Country: Egypt/ living in Germany**

**University: Siegen.**

**Specialization: NLP**

**Problem Description:**

As mentioned, on the website of the internship, document / Text classification is one of the important applications in supervised machine learning (ML). Many of news websites try to recommend similar news to the reader. The process of recommendation depends on the category of the news. News should be classified and recommended to the users based on that. The challenge is to build a good ML system to predict the category of the online news with high accuracy.

For Example – New York Times are using topic models to boost their user – article recommendation engines. Various professionals are using topic models for recruitment industries where they aim to extract latent features of job descriptions and map them to suitable candidates. They are being used to organize large datasets of emails and customer reviews and use social media profiles.

**Data understanding**

The data that I am using is the famous "20 Newsgroups" dataset.

The data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups.

**What type of data you have got for analysis?** the data is in textual form.

**What are the problems with the data?** The data is unstructured, the data should be read into a notebook using python and some operations are done on the data to extract the target and the articles from each file. The files containing the articles have no extension but using python, it is easy to read them. Also, the content of each file is messy and contains unuseful information that should be removed using some cleansing techniques.

**What approaches you are trying to apply to your data set to overcome problems?** I will Clean the data by removing the stop words, the URLs, the special characters, metadata and digits.

Github Repo link

**What type of data is given the analysis** **-** .xlsx type of data

**Data shape –** 3424 by 69

**Variable types –** categorical variables are 67 and the numerical variables are 2

**What problems are with the data:**

* **Number of NA values –** 0
* **Number of duplicates -** 0
* **Outliers –** There are outliers in the numerical columns; Dex\_Freq\_During\_Rx and Count\_Of\_Risk
* **Skewed data –** The numerical columns have skewed data.
* **Class imbalance in the target variable –** There’s a slight imbalance

**Solutions to the problems**

1. **Outliers**

* Dropping the outliers by using the Inter-Quartile-Range method

1. **Skewed data**

* Dropping the outliers (worked perfectly).
* Log transformation of the numerical features
* Applying the min-max to the numerical features

1. **Class Imbalance**

* Oversample the minority class
* Downsample the majority class

**Vocabularies**

1. **Drug persistency –** The act of continuing the treatment for the prescribed duration.
2. **Adherence –** The extent to which a patient acts in accordance with the prescribed interval and dose of a dosing regimen.
3. **IDN -** a network of healthcare providers and facilities within a specific geographic region that offers a full range of healthcare services. An IDN is often designed to offer a full spectrum of care inclusive of primary care physicians, specialists, general acute care (i.e. inpatient services), and home health services.
4. **NTM –** Non-Tuberculosis Mycobacteria
5. **Rx –** Medical prescription
6. **DEXA Scan –** Also known as bone density scan. Used to measure calcium and other minerals in the bone.
7. **Fragility fracture –** a fracture resulting from a fall.
8. **Glucocorticoid –** Are steroid hormones used for the treatment of inflammation, autoimmune diseases and cancer.
9. **Injectable drugs –** Drugs injected to users
10. **Comorbidity –** condition of having two or more diseases at the same time.
11. **Concomitancy –** Existing or occurring together. Concomitant medication are other prescription drugs, over the counter drugs or dietary supplements that are taken other the drug under investigation.
12. **T score –** Measures how much a bone density is higher or lower than that of a healthy 30-year-old adult.

Data shape – 3424 rows x 69 columns

Categorical variables – 67

Numerical variables - 2

Problems with the data:

Number of missing values – 0

Outliers – There are outliers in the numerical columns; Dex\_Freq\_During\_Rx and Count\_Of\_Risk

Skewed data – The numerical columns have positively skewed data.

**Data Understanding:**

The healthcare dataset is considerble size data having 69 columns altogether. The label is the

Persistency\_Flag which is a binary data having value as True or False depending on the other

features. The first column being the unique id of the patient is of no use for us as this is not going to

help us in training the model. Hence the first thing we do is drop that column. We have analyzed the

dataset and upon analysis we found that there are only few numerical data column and rest are

either binary or string value.

**Exploratory Data Analysis :**

We analyzed and dataset and found that there are many binary data having “Y” and “N” values. We

mapped each if the data to 1 and 0 respectively. Also, we find no null values in the dataset and

hence there was no need of handling it. We checked the numerical columns and found that one of

the feature is having some outliears and we handled it using log transformation.

**Github Link:** https://github.com/aniyaz/Healthcare-DataScience2021

**Submitted by: *Aman Niyaz***

**Date: *12h April 2021***

Name: NLP: Document Classification

Report date: 22/06/2022

Internship Batch: [LISUM10: 30](https://canvas.instructure.com/courses/4851447)

Version: 1.0

Data intake by: Amr Elbana

Data intake reviewer: All Members

Data storage location: UCI

**Tabular data details:**

|  |  |
| --- | --- |
| **Total number of observations** | 19997 |
| **Total number of files** | 20 |
| **Total number of features** | 2 |
| **Base format of the file** | .txt |
| **Size of the data** | 45 Mb |