Data Science Case Document

Identifying online patient conversations

# Rules and Regulations

* Every Individual will solve the challenge separately
* The Training and Testing Data are available at this location -> [link](https://drive.google.com/drive/folders/0B6CR4opOGS2WYVFLWHozZ2ptX00?usp=sharing)
* Time Limit for sending your submissions will be 3.5 hours (10 am – 1.30 pm)
* Time Limit for making PowerPoint Presentations will be one hour (1.30 pm – 2.30 pm)
* Please follow the flow and the structure of the presentation as per the template attached



* Please Submit your files in the following format
  + Evaluation file (test outputs) -> FirstName\_LastName\_dateofbirth.csv (date of birth in dd/mm/yy format)
  + PowerPoint Presentation -> FirstName\_LastName\_dateofbirth.ppt
* Please mail your submission file to [vinit.gela@zs.com](mailto:vinit.gela@zs.com) by 1.30 PM and your presentation file to the same email address by 2.30 PM
* Details about the format of the submission file and data are given below in the document. Please read it carefully

# Problem Introduction

ZS Data Science team collaborates with the Social Listening team for automating the process of gaining insights from social media conversations. Pharma clients are really interested to know what insights or recommendations the team can give them using such non-formal digital mediums. Patients are one of the key stakeholders for the clients, and hence insights around what patients are speaking online becomes quite important. In order to do so, the Social Listening team first identifies the patient conversations from the rest of the conversations posted by others. If we take the case of heart failure, the team analyses around 60-70K “heart failure” related conversations (per month) fetched from forums, twitter, Facebook, blogs and news sites to segregate the patient conversations. This takes a huge manual effort, which requires a person to work for 8 hrs. Per day for two weeks to identify the patient conversations. The ZS Advanced Data Sciences team want to help them out by predicting such conversations using machine learning and automate the process of identifying patient conversations.

# Problem Description

The Social Listening team has to manually validate heart failure related conversations fetched from the social listening tool which scans twitter, Facebook, forums, blogs etc. Such conversations are posted by multiple stakeholders like patients, doctors, media houses, general public, etc. The team needs to identify the patient conversations, so as to dig deeper into them and identify the patient needs. The data science team wants to automate this process by building intelligent algorithms to predict patient conversations.

Build an Intelligent pipeline that can segregate patient conversations from the rest of the group given historically tagged patient data – The Social Listening team has manually tagged a sample of data with flag 1 for patient conversations and flag 0 for non-patient conversations

The conversations that are non-patient can be like

* Medical Advise by HCPs
* Generic information by healthcare organizations
* News article or statement released by press etc.

The conversations by patients will mostly talk about their problems, their health conditions, their experiences and questions seeking advice.

The problem is to build an algorithmic solution to predict patient conversations to speed the process of patient insight generation

# Data Description

The Social Listening team has previously manually validated data for heart failure conversations with patient conversations tagged as shown below in the table

The headers of the training file are

|  |  |
| --- | --- |
| **Header** | **Header Description** |
| **Source** | **Type of Social Media Post** |
| **Host** | **Domain of Social Media Post** |
| **Link** | **Complete URL of post** |
| **Date** | **Date of Post** |
| **Time** | **time Stamp of Post in Eastern Time** |
| **Time(GMT)** | **time Stamp of Post in GMT** |
| **Title** | **Title of the Post** |
| **TRANS\_CONV\_TEXT** | **Actual Text Conversation of the Post** |
| **Patient\_Tag** | **Patient Flag (1= Patient, 0=Non-Patient)** |

Sample Conversations are as below

|  |  |
| --- | --- |
| TRANS\_CONV\_TEXT | Patient Flag(1 = Patient, 0 = Others) |
| I was admitted to the hospital for 5 weeks due to a sudden cardiac arrest. Drug X was used continuously during that period which led to a drastic weight loss and weakness | 1 |
| One of the important benefits of drug Y is the cardiovascular benefits along with control of insulin for type 2 diabetes | 0 |

# Desired Solution

The Team would like to build an algorithm where they can ingest the social data and get the patient tags - 1 if patient and 0 if not a patient.

If possible the team would also like to know the confidence of prediction and which features led to the prediction

For the Testing, Please send your outputs (Predictions of 1 and 0 on the test dataset) in a csv file as header “Output” as shown below. The name of the submission file should be FirstName\_LastName\_dateofbirth.csv. Email this submission file to [vinit.gela@zs.com](mailto:vinit.gela@zs.com) by 1.30 PM

|  |  |
| --- | --- |
| Index | Output |
| 0 | 0 |
| 1 | 1 |
| 2 | 0 |