DTI5125[EG] Data Science Applications Project Report  
Personalised recommendations based on user’s tweets

short line

Ali Elshenawy

Ahmed Elemary

Hassan Hefny

Sarah Ibrahim

Contents

[Problem formulation 1](#_Toc110201278)

[Data Preparation 2](#_Toc110201279)

[Integration 2](#_Toc110201280)

[Cleaning 3](#_Toc110201281)

[Text feature engineering: 3](#_Toc110201282)

[Classification models 4](#_Toc110201283)

[Clustering model 5](#_Toc110201284)

[Error analysis 9](#_Toc110201285)

[1. Docs Avg length 9](#_Toc110201286)

[2. Fields Distribution 9](#_Toc110201287)

[3. Words frequency: 10](#_Toc110201288)

[Test case 11](#_Toc110201289)

[Future Work: 14](#_Toc110201290)

# Problem formulation

The project aims to identify users' interests by classifying their tweets into categories; eg: sports, entertainment, politics etc. And based on that, recommend other tweets that the users are more likely interested in.

Project outline:

Classifier

User tweets

45% sports

30% Politics

15% financial

10% Entertain

0% Medical

Other sports tweets

Clastring

Other politics tweets

Recommended tweets

# Data Preparation

### Integration

1. Tweets classification data set from Kaggle platform [↗](https://www.kaggle.com/datasets/pradeeptrical/text-tweet-classification) containing labelled data of four categories: (Sports, politics, entertainment, and medicine). The data set contains 1162 rows, 2 columns
2. The second data set contains financial tweets for sentiment analysis purposes [↗](https://www.kaggle.com/datasets/vivekrathi055/sentiment-analysis-on-financial-tweets). It contains 8 columns, only the text column will be used to add the financial category to the first data set. The data set contains 28442 rows. A random sample of 300 rows will be chosen to keep the final data set balanced.

Final data frame:



### Cleaning

Rejex was used to:

1. Remove user name from the tweet
2. Remove tweets URLs
3. Make sure that only English characters and numbers are included

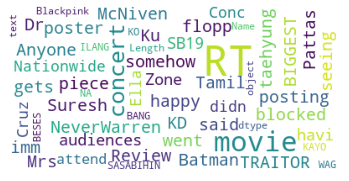
# Text feature engineering:

* Bag of words transformation was applied to the text to use it with the Kmeans classifier.
* PCA feature reduction to two features was applied on the BOW for training and visualization

The following figures show the word cloud of each category:



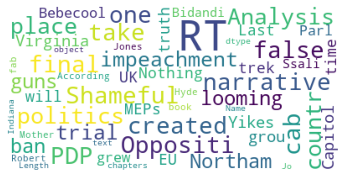
Sports



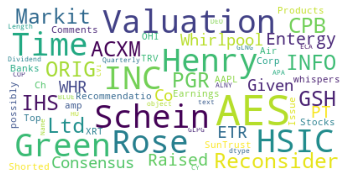
Entertainment



Medical



Politics

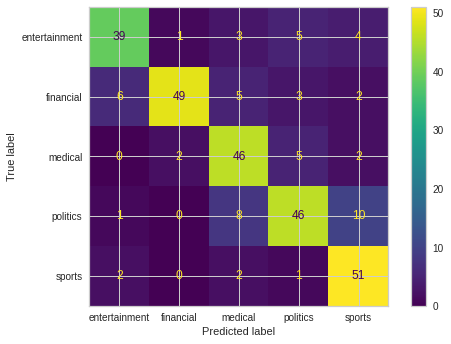
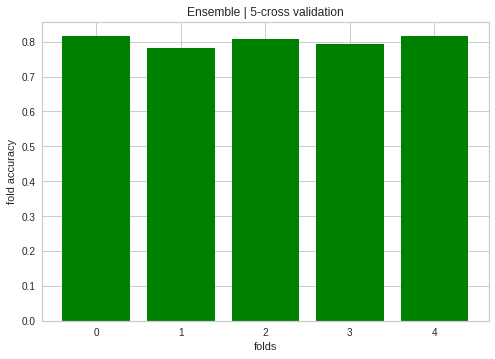


Financial

# Classification models

* We applied heterogeneous ensemble learning on the following classifiers and calculated the final label results using soft voting

1. Support vector Machine
2. Naive Bayes classifier
3. Random Forest.

* We applied 5-fold cross validation on this model while training. Average accuracy 80.32%
* we tested the final classification model on the test data. Accuracy of the model on the test set = 78.83%

# Clustering model

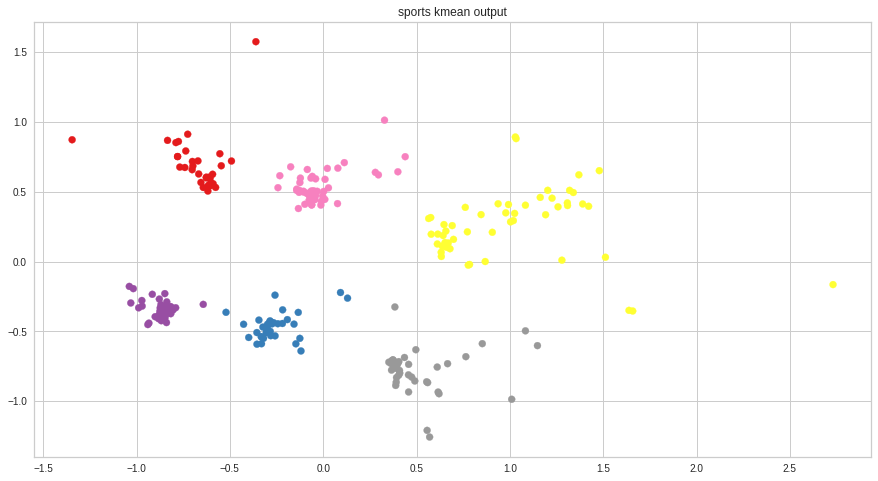
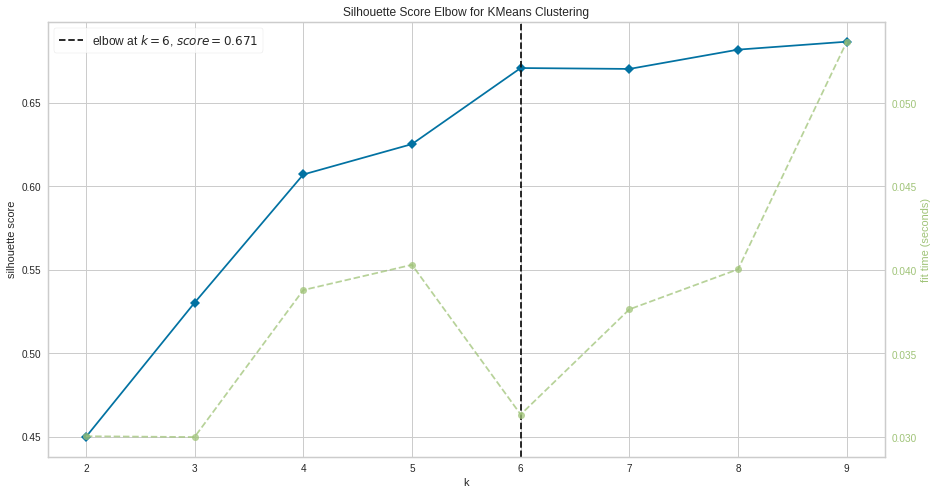
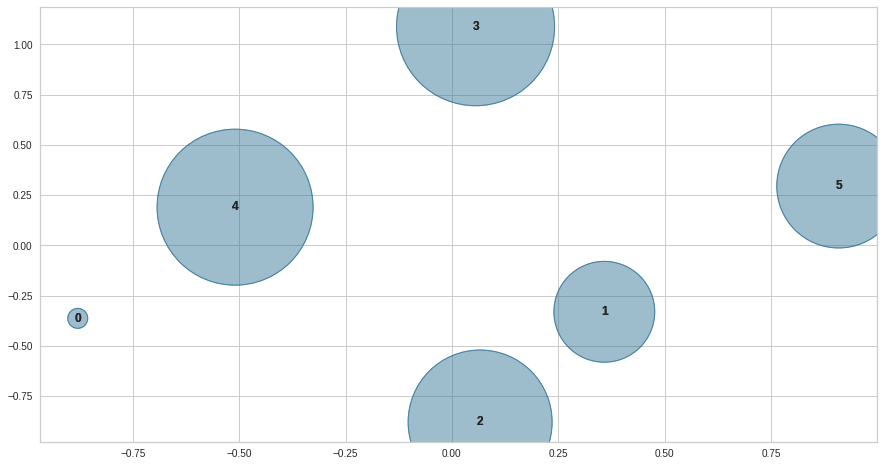
If two users are interested in sports, that does not mean that they are interested in the same topics of sport, that’s why we need to cluster each field to determine which more specific tweets the user prefers.

#### Steps:

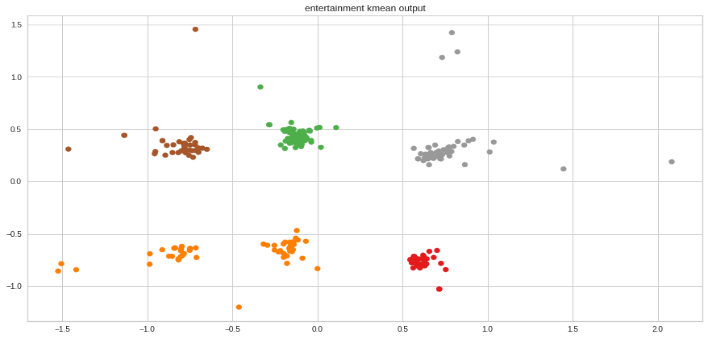
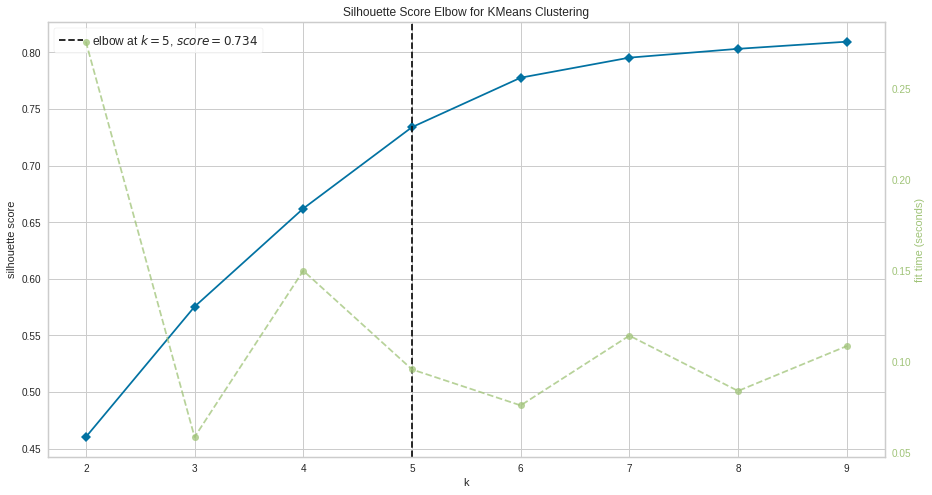
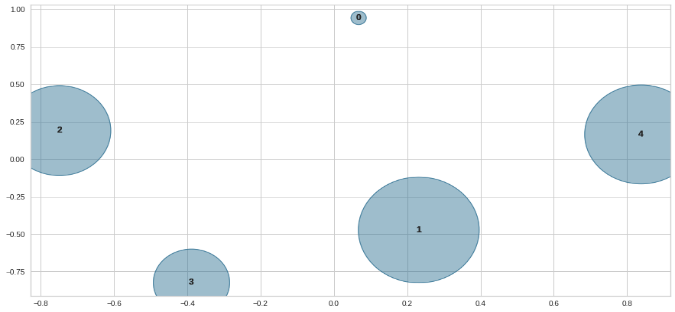
1. Filter the dataset to get all of the rows that contain sports as a type column in the data set.
2. Transform the data from categorical data to numeric data so we use BOW to transform.
3. use PCA as a technique of unsupervised dimensionality reduction to minimize our features to plot the fields clusters

The following figures show each field Silhouette & WSS, clusters, and distances between clusters:

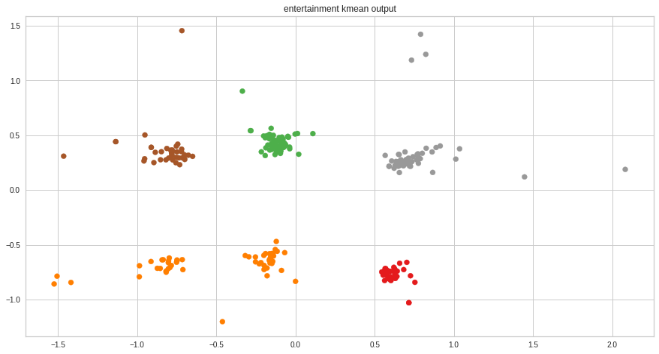
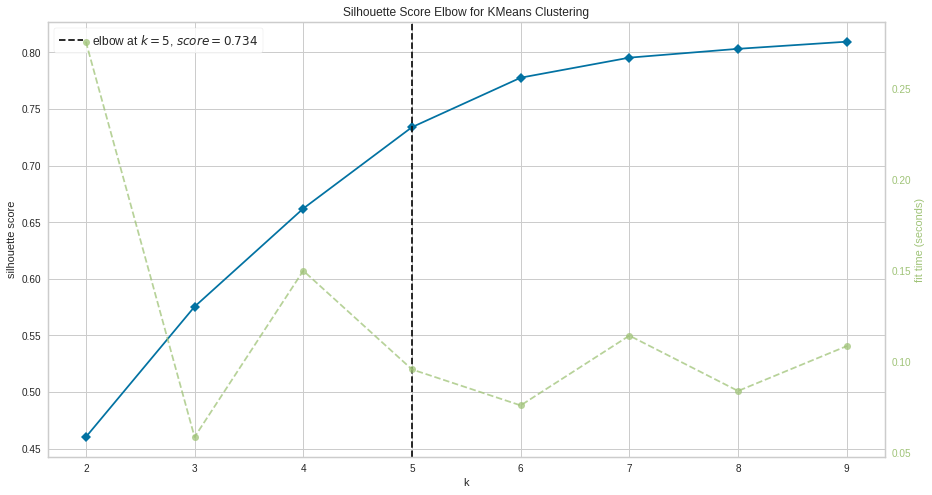
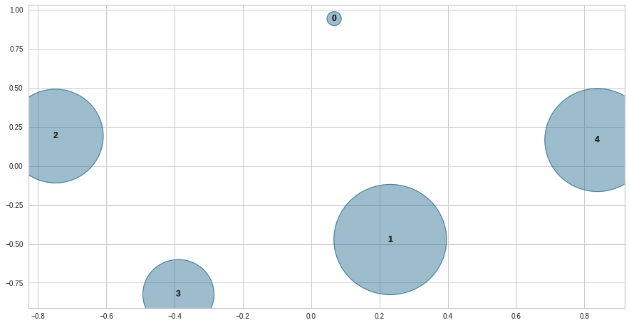
##### Sports:



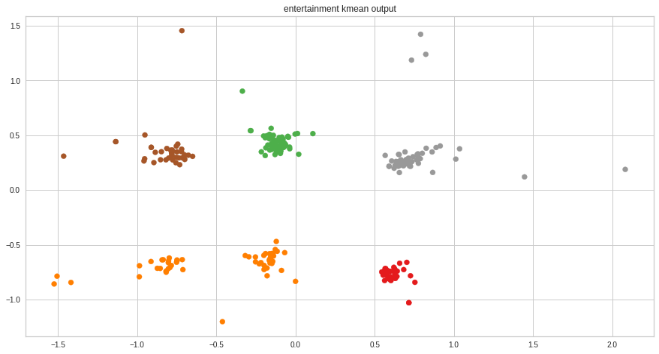
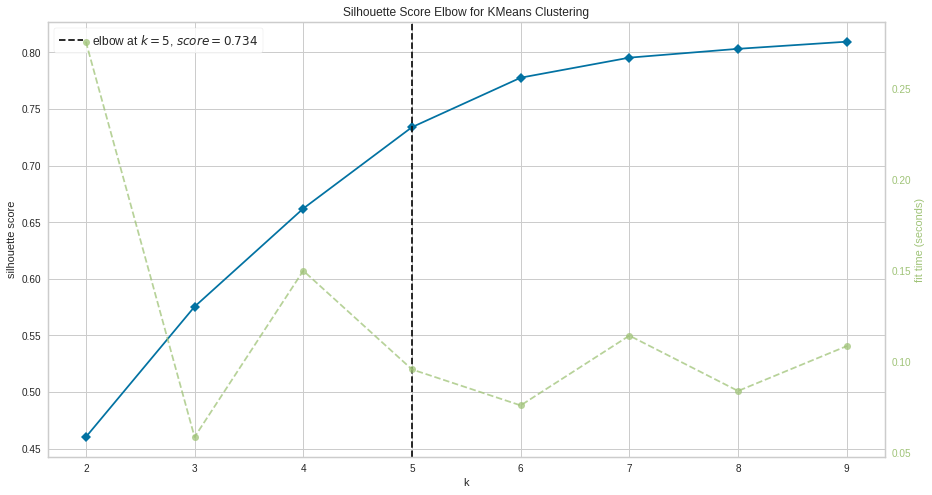
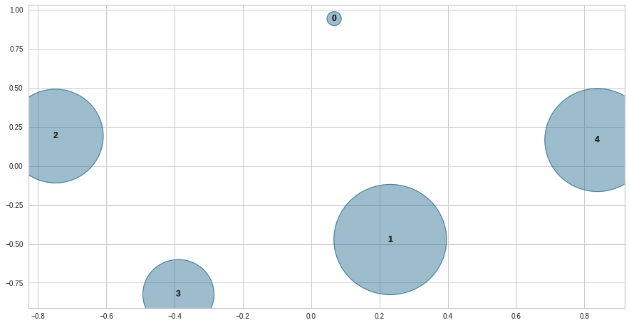
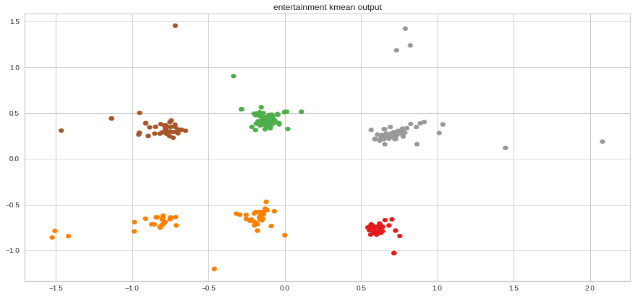
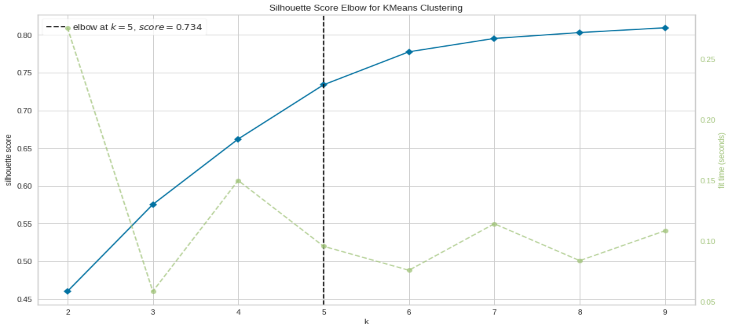
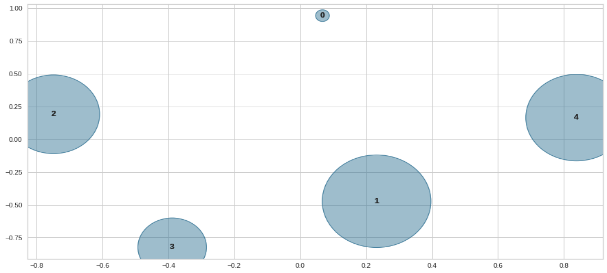
##### Entertainment:



##### Medical:



##### Politics:



##### Financial:

# Error analysis

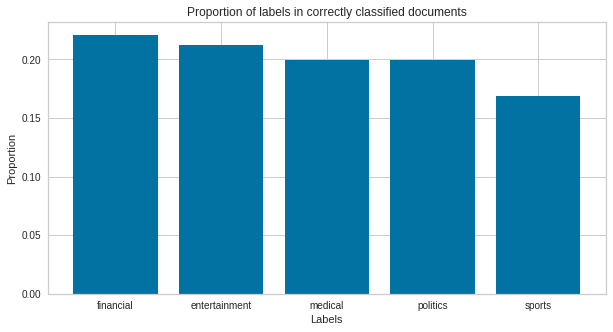
In this project three approaches of error analysis are used:

* Documents average length in correct predictions VS Documents average length in wrong predictions.
* Fields Distribution in correct predictions VS Fields Distribution in wrong predictions.
* Finally, most frequent words that could be the reason why model had misclassification.

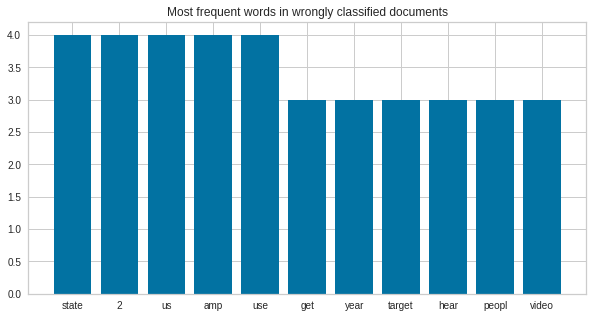
### Docs Avg length

documents length in both wrongly and correctly predicted samples are almost the same as the have nearly the same average value (64). So, there is no problem with documents length

### Fields Distribution

* In wrongly predicted samples, about 55% of them are of the “Medical & Entertainment” fields
* In correctly predicted sample this is not the case as samples are uniformly distributed with respect to fields.
* This means that bias towards “Medical & Entertainment” are not possibly just a chance and could be vital part of the problem.

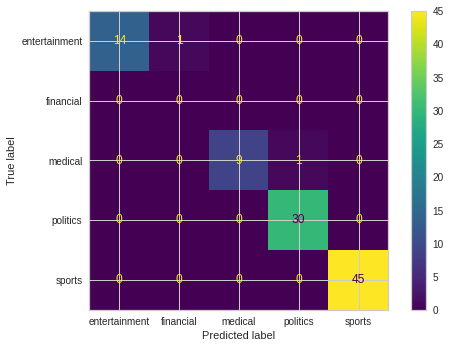
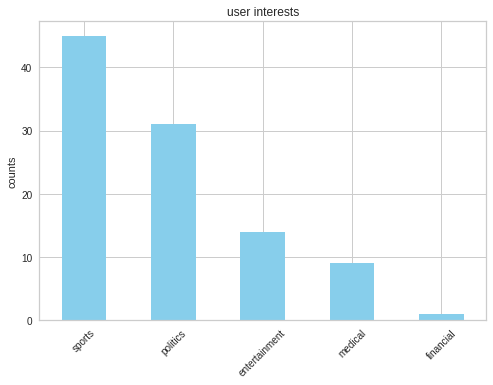
### Words frequency:



* In the right-bottom fig, we can see the most frequent words in the wrongly classified samples.
* Most of these words don’t even exist in correct prediction samples.
* More interesting observation that in correctly-classified docs many of these words don’t exist in wrong predicted-docs, despite the fact that some of them have frequency > 20, while the most frequent word in the wrong-docs has only frequency = 4

# Test case

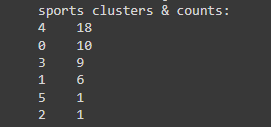
* To test the overall model, we extracted 100 random tweets from the data set with the following parts:(45% sports, 30% politics, 15% entertainment, 10% medical, 0%financial)
* First, the classifier model should successfully determine the user's top interests:

Classifier output:

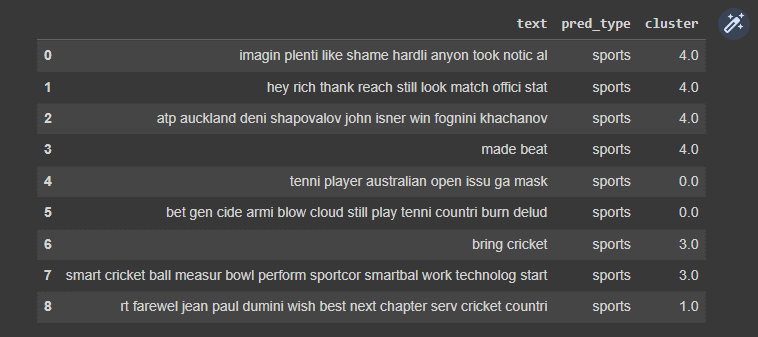
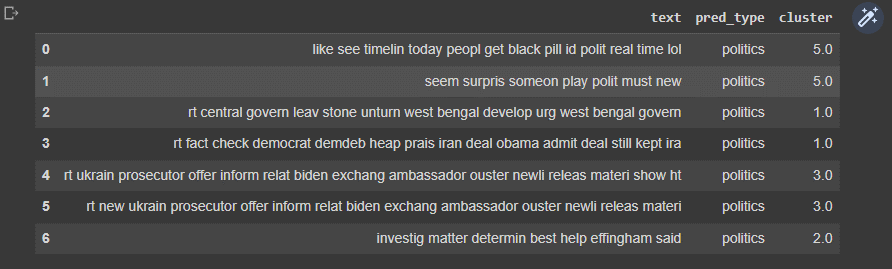
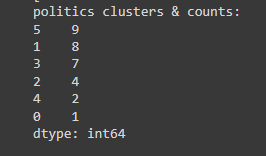
So, the test user's top interests are sports (45%) and politics (31%).

* The next step is to Find out the user prefers which sports and politics tweets.
* Recommend similar tweets with the same percentages.

Example:

n(4) =

n(4) =



# Future Work:

* Take into consideration, depending historical models such as LSTM, we provide tweets history, in our data, so that we become able to if a user is still interested in a field he no longer does
* Make the model ready for Deployment specially for twitter platform
* May be apply association rules to which fields a user may be interested in, in the same time

short dash