

# Case Study Project

Name: Anirudh Sundaresan

## **Background Information:**

- Social media is not just a platform where people talk to each other, It has become a medium where people express their interests, share their views and displeasures.
- Internet is a major source of spreading terrorism through speeches and videos. It has become necessary to analyze this information to avoid any form of hate crime or violence.

## **Data Collection:**

**Data Origin-** Kaggle. We took half the data of pro-ISIS fanboys' Tweet and the other half from normal Tweets ( having a combination of negative, positive, or neutral sentiment). This will ensure that we have good variety of data to train our model. Total size of the dataset is about 10,000 records.

**Key Attributes-** Type, Tweets

### **Data Quality-**

- Not in readable format for Naïve Bayes to work on.
- Categorical variables are not factored.
- Cases, numbers, punctuation, filler/stop words, white spaces need to be processed so that Naïve Bayes classifier can work on the dataset.

## **Problem Statement:**

- Our proposed project aims at text mining to check prominence of a keyword commonly used in terrorist activities. So, this system can be used to point out users spreading hate crime.

## **Objective:**

- To clean the data, analyse and visualise it.
- Predict the comments showing signs of hate crime.