## **Capstone Project Domain #11 ( Sentiment Analysis in Twitter )**

**Introduction**

Social media has opened the floodgates of customer opinions and it is now free-flowing in mammoth proportions for businesses to analyze. Today, using machine learning companies are able to extract these opinions in the form of text or audio and then analyze the emotions behind them on an unprecedented scale.

Understanding people’s emotions is essential for businesses since customers are able to express their thoughts and feelings more openly than ever before.

In this regard, sentimental polarity prediction in social media (Classification of a tweet as negative or positive or neutral) is highly valuable for certain institutions, organizations.

“When captured electronically, customer sentiment — expressions beyond facts, that convey mood, opinion, and emotion — carries immense business value. We’re talking the voice of the customer, and of the prospect, patient, voter, and opinion leader.”

— Seth Grimes

**Problem Definition**

Tweet text along with other features has been extracted from different from different sources (domain) using APIs. ( We will not extract using API , Live data will be provided)

Each row of the dataset contains sentiment code (negative, positive and neutral embedded in Twit-id column. The task is to predict whether a tweet contains positive, negative, or neutral sentiment. This is a supervised learning task where given a text string.

**Hardware & Software planning:**

|  |  |  |
| --- | --- | --- |
| Hardware / Software | Details of Resource | Remarks |
| Laptops | Windows OS, At least 8 GB RAM and i5 or above processor |  |
| Software | ANACONDA with Python 3.7.3  Jupyter Notebook 6.0.3  Google Colab |  |

**Dataset**

Rows 38561

Column 10

Column description

1. Tweet\_id
2. SourceDataBase
3. OS
4. Tweet-Class\_category-Code
5. Tweet\_source
6. Tweeted-By
7. Retweet\_count
8. Tweet
9. Date Created
10. Country

**Important Tasks**

1. Data import and standard data processing and cleaning
2. Tweets Preprocessing and Cleaning
3. Visualization
4. Extracting Features from Tweets
5. Model Building: Sentiment Analysis – Various ML-Algorithm will be used for better accuracy)
6. Model Fine-tuning
7. Final Data Preparation

**Important Python Libraries for text processing**

import re

import nltk

nltk.download('wordnet')

nltk.download('stopwords')

from nltk.corpus import stopwords

from termcolor import colored

from nltk.stem import PorterStemmer

from nltk.stem import WordNetLemmatizer

from wordcloud import WordCloud, STOPWORDS

##### from sklearn.feature\_extraction.text import TfidfVectorizer ( bag of words approach - TF-IDF)