# *Sentiment Analysis on IMDB movie reviews using NLP*

**Chandramouli Yalamanchili**
**DSC680 - T302 Applied Data Science (2215-1)**
**https://chandu85.github.io/data-science**

**Proposal Document** - https://github.com/chandu85/data-science/blob/main/Project%203%20-%20IMDB%20Movie%20Review%20Sentiment%20Analysis/Documentation/Project%203_Proposal.pdf

## Project details

**Project Domain:**

I chose to work on NLP, Natural Language Processing analytics for my 3rd project, especially how it can be used to measure human emotions with respect to a particular topic. For this project I chose to work with IMDB movie reviews.

**Project Abstract:**

Natural Language Processing (NLP) helps the machines to understand the human language, it is a component of artificial intelligence. NLP helps machines to understand the text data that can come from many sources and is unstructured in nature. Great progress has been made in last few decades in NLP domain mainly because of increased computing capabilities, we see NLP being implemented in lot of applications around us through several smart home devices and chat bots.

Sentiment analysis is an NLP technique that is used to determine the emotion attached to the textual data that generally represents the comments from humans. Sentiment analysis is playing a key role in several enterprises already by enabling the automation in the area of analyzing the customer reviews and feedbacks there by attaining the customer feedback much quicker and being able to make decisions for future products faster while meeting the customer demands.

Through this project I will explore different NLP models and text classification models to analyze the IMDB movie reviews dataset and come up with the sentiment for each review comment to indicate whether it was a positive comment or a negative one.

## Week 2 check-in

Any surprises from your domain from these data?

- I have not come across any challenges or surprises with respect to the domain.
- I am spending time to understand how NLP is helping in different business sectors. I am also spending time to learn more about how sentiment analysis is helping different business sectors.
- It seems like there is a lot of progress made in both of these domains in last couple of decades. I am trying to review several referrals I have gathered on these topics.
- I am hoping to have my literature review complete by end of tomorrow, 5/31 to be included in my final project report.

The dataset is what you thought it was?

- Yes, data was what I thought it would be, no challenges so far with respect to dataset itself.

- Dataset has total of 50,000 IMDB movie reviews, it has two columns one with the review text and the other column with the class indicating whether it is a positive or negative review.
- The dataset is very well balanced with 50% positive and 50% negative reviews.
- Overall no major concerns or issues in working with the dataset itself.

Have you had to adjust your approach or research questions?

- Not really, at least at this time I am still going ahead as per my plan at this time.
- I did run into few issues in trying to get tensor flow installed and configured on my laptop, but I still have the same goals as I have started with.
- I have just started with EDA, will be finishing up with EDA and modeling by tomorrow, 5/31.
- At this time, I am still going ahead with the same plan and with the same research questions I started with.

Is your method working?

- I am planning to use tensor flow and nltk packages to do the EDA and data preparation mainly to clean up the data by eliminating the punctuation and also to apply tokenization techniques to get the text ready to be used for modeling.
- At this time, I am exploring multiple modeling techniques like nltk, scikit-learn based standard modeling algorithms vs. deep learning techniques like RNNs, BERT, etc to see which option works best for me and give me better performance.
- If time permits, I will build multiple models and compare the performance among them.

What challenges are you having?

- I did run into a major roadblock yesterday while trying to get tensorflow-text setup for using in my anaconda based Jupyter notebook. It seems like tensorflow-text is not available yet for anaconda. I have managed to switch to get tensorflow-text setup through PIP and use visual studio code for building my project.
- Once I have all of the necessary package's setup, things seem to be going smooth for me at this time.
- At this time, I don't have any critical issues, quite a few things are new to me as I have not worked with tensor flow and NLP before so it's been really interesting so far.
- I am learning quite a few new things, progressing well and hoping to have everything completed for this project by 05/31.