Thumbs Up or Thumbs Down Analysis:

Using Natural Language Processing to Predict Movie Title Ratings

Peter Adam Aboud   
*Department of Computer Science*  
*Ryerson University*Toronto, Canada  
adam.aboud@ryerson.ca

Maria Galsim  
*Department of Computer Science*  
*Ryerson University*Toronto, Canada  
maria.galsim@ryerson.ca

# Introduction and Motivation

Internet streaming services have exploded in popularity in recent years, becoming an extension of cable television or a direct competitor itself. This has led to a proportional increase in demand for content that appeals directly to specific users and content recommended to cater to the personalized experience. Classifying and scoring content have become a pivotal part of the business model for streaming services such as Netflix, Amazon Prime Video, Hulu, etc. Being able to accurately recommend content to users is an emergent problem for these companies and there is an emphasis to find the optimal solution. This application of machine learning will endeavour to show that Natural Language Processing can be used to accurately predict title rating score and can be used as another tool in the recommendation process.

# Methods

## Data Cleaning and Processing

The following datasets will be considered and used during the course of our project. In order to render them into a usable format, we will use the python library ***pandas*** to adjust them to our needs.

* https://www.kaggle.com/shivamb/netflix-shows
* https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/data
* https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
* https://www.kaggle.com/lakshmi25npathi/sentiment-analysis-of-imdb-movie-reviews

## Text Normalization and Pre-Processing

The first step in Natural Language Processing is the normalization and pre-processing of our input data. This includes:

* Tokenization
* Numericalization
* Syntactic preprocessing

There are a number of useful python libraries that will help with this, including ***fastai***.

## Sentiment Analysis

### Tier 1:

* We will be taking in a ‘bag of words’ that have been rated positive or negative and then extracting features with the help of NLP toolkit for Python.
* After the number of words are segregated into good or bad and being counter for each type, these will become the two features.
* This model will be evaluated based on it’s ability to predict a rating from 1-10, where:
  + Rating > 5 is positive
  + Rating < 5 is negative

### Tier 2:

* Building on the ‘bag of words’ technique:
* We will be converting each input text sample into a word vector, which will be representative of the occurrences of specific words.
* Apply logistic regression
* This model will be evaluated based on its ability to predict a rating from 1-10, where:
  + Rating > 8 is positive
  + 8 > Rating > 6 is somewhat positive
  + 6 > Rating > 4 is neutral
  + 4 > Rating > 2 is somewhat negative
  + 2 > Rating >0 is negative

### (POSSIBLE) Tier 3:

* Apply more advanced deep-learning techniques to the problem of multi-class text classification
* Use techniques such as: discriminative fine tuning, Universal Language Models, etc. in order to improve the accuracy of our model

## Failure Analysis and Re-Engineering

* Conduct a review of the efficacy of each model tier and evaluate accordingly
* Look into where the failures are with analysis the review
* An example could be that a review is worded positive but actually has a negative meaning
* Re engineering may include:
  + Including additional pre-processing techniques
  + Multiple layers of neural networks
  + Applying new language models

## Evaluation Metrics

* We will start by looking at the model from a regression type
* We have not fully thought out what the other evaluation metrics will be

# Intended Experiments

We intend to begin our Sentiment Analysis using the simplest form of the ‘Bag of Words’

* Positive and negative words will be counted and used as features to train a linear regression model in order to predict the output rating.
* This model will be evaluated with how accurate the predictions were even with this simplest model

Taking into the account the results of the Tier 1 model, we will be adding the vector model and then evaluating if this change produces more accurate results and granularity

# Planning and Milestones

## Milestone 1

* By October 2, 2020
* Research of datasets and submission of project proposal
* To be completed by: Adam and Maria

## Milestone 2

* By mid-October
* Data pre-processing, clean up and feature extraction
* To be completed by: Adam and Maria

## Milestone 3

* End of October
* Sentiment Analysis design and training
* To be completed by: Adam and Maria

## Milestone 4

* By end of November
* Testing and failure analysis
* To be completed by: Adam and Maria

## Milestone 5

* Timeframe
* Re-engineering and re-training of our model
* To be completed by: Adam and Maria

## Milestone 6

* By December 15, 2020
* Complete the final report and prepare the 5-minute video presentation
* To be completed by: Adam and Maria

##### References

1. https://www.kaggle.com/shivamb/netflix-shows
2. https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/data
3. https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
4. https://www.kaggle.com/lakshmi25npathi/sentiment-analysis-of-imdb-movie-reviews
5. Maryanchyk, Ivan, Are Ratings Informative Signals? The Analysis of the Netflix Data (October 2008). NET Institute Working Paper No. 08-22, Available at SSRN: <https://ssrn.com/abstract=1286307> or [http://dx.doi.org/10.2139/ssrn.1286307](https://dx.doi.org/10.2139/ssrn.1286307)
6. Roberti, Maximilien, Fastai with 🤗Transformers (BERT, RoBERTa, XLNet, XLM, DistilBERT) https://towardsdatascience.com/fastai-with-transformers-bert-roberta-xlnet-xlm-distilbert-4f41ee18ecb2
7. Amatriain X., Pujol J.M., Oliver N. (2009) I Like It... I Like It Not: Evaluating User Ratings Noise in Recommender Systems. In: Houben GJ., McCalla G., Pianesi F., Zancanaro M. (eds) User Modeling, Adaptation, and Personalization. UMAP 2009. Lecture Notes in Computer Science, vol 5535. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-02247-0\_24