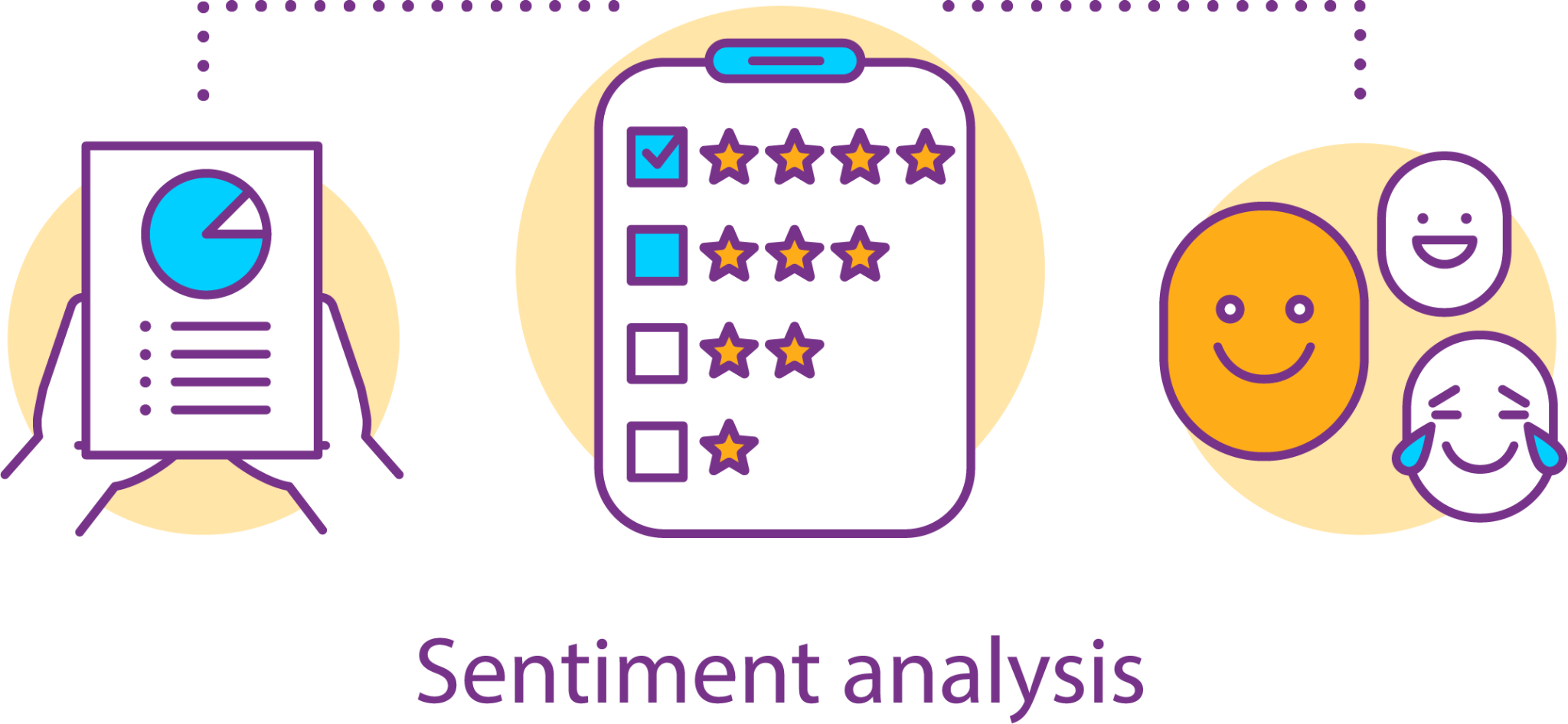
**Sentiment Analysis of Textual Content using Natural Language Processing**

****

****

Submitted By:

B. Abhinav : AM.EN.U4CSE20216

C. Pranav Vamsi Krishna : AM.EN.U4CSE20220

G. Praveen Naidu : AM.EN.U4CSE20224

G. Nandan Chakravarthi : AM.EN.U4CSE20226

**M. Snehashish Reddy : AM.EN.U4CSE20246**

O. Hitesh : AM.EN.U4CSE20252

ABSTRACT

Our Team will be constructing a program that can identify the emotion behind a piece of text, written by a person. The text could be anything. It could be a tweet, or it could be an excerpt from a novel. In our case, it is going to be a movie review on Rotten Tomatoes.

The model should be capable of classifying an unpublished movie review by a critic, based on a corpus of reviews on Rotten Tomatoes all of which have been classified as negative, somewhat negative, neutral, somewhat positive, and positive, as measured by the movie’s Tomatometer. The classification accuracy would be determined by comparing the percentage of the model’s predictions with the test data, labeling each review with:

0 - Negative

1- Somewhat Negative

2 - Neutral

3 - Somewhat Positive

4 - Positive

ASSUMPTIONS

* Most reviews in a huge dataset would have similar words associated with their label. Ex: Hate -> 0, Dislike -> 1, Ok/Okay -> 2, Decent/Good -> 3, Amazing -> 4.
* Sentence Imbalance could be used to improve accuracy in the case of Sarcastic Reviews.
* Algorithms like KNN might have to be employed here to broadly classify the reviews into their labels.

1. Introduction

**Motivation:**

The problem is not meant as just a learning exercise. The objective is to transition the research we do in this project into a paper soon. Natural Language Processing has been a tough nut to crack and has very standard methods and recurrent procedures. We want to explore new techniques across the internet and research papers to make it simpler to process and create. If in the case we discover something that was not before, we will definitely be authoring a research paper on the topic. NLP will also give us a good idea of how machines' process works behind the scenes on applications like Google Translate. We are training this model with purely the intent of making it a starting point for further advanced NLP models.

**Benefits of Solution:**

The data classified by this model will help identify the sentiments of certain movies in the current critic space. For ex: In a particular year, action movies might be more popular than fantasy movies in the eyes of the critics. Actively training this model will help predict the sentiments of movies which then distributors can utilize to help maximize their profits. This will also help the producers help plan movie release dates according to the sentiments. These are just a few of the ways the data obtained from this model can be used.

**Solution Use:**

We expect a long lifetime for this model. The words in the English language have remained consistent. With properly supervised learning, this model can remain in use for an exceptionally long time. Even longer as a starting point for training advanced NLP models. The solution is mainly meant for corporate use, thus every year larger and larger datasets would be required to professionally train the model which could be the only drawback to this model but, at the same time, the accuracy will not be dropping over time for the movie reviews. The model could easily be transitioned over to Twitter or similar social media platforms to identify toxic posts and help comply with government censorship on certain topics. One of the most famous tools in the market for content checking, Grammarly, also works on similar principles

2. Dataset Finalization

Being inspired by a Kaggle Competition, we will be using the exact same datasets that were originally collected by Pang and Lee. The Rotten Tomatoes movie review dataset is a corpus of movie reviews used for sentiment analysis inspired by the works of [Richard Socher](https://www.socher.org/).

The data itself can be obtained [here](https://www.kaggle.com/competitions/sentiment-analysis-on-movie-reviews/data).

There are three datasets. One for the sample submission, one to train the model, and one to test the model’s accuracy. If needed, the training dataset and the testing dataset will both be partitioned according to the requirements which will promptly be updated here.

1. The dataset is comprised of tab-separated files with phrases from the Rotten Tomatoes dataset. The train/test split has been preserved for the purposes of benchmarking, but the sentences have been shuffled from their original order.
2. Each Sentence has been parsed into many phrases by the Stanford parser. Each phrase has a PhraseId. Each sentence has a SentenceId. Phrases that are repeated (such as short/common words) are only included once in the data.
3. The data has previously been used to research [Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank](https://nlp.stanford.edu/sentiment/)

3. Project Updates (Classes and Datasets)

Though initially inspired by a Kaggle competition:

* The no. of classes of the datasets have been reduced from multiple classes (0, 1, 2, 3, 4) to Binary (0,1) and thus, the model will now be trained for binary classification of sentiments in movie reviews
* The old Kaggle dataset has been replaced in favor of 5 well-classified datasets from various sources and 1 raw training dataset from Stanford

**Final Datasets:**

1. [IMDB Dataset of 50K Movie Reviews](https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews)
2. [ACL IMDB Raw Corpus (Two Datasets)](https://ai.stanford.edu/~amaas/data/sentiment/)
3. [SST Binary v2](https://huggingface.co/datasets/sst2)
4. [Rotten Tomatoes Scraped Reviews (Two Datasets)](https://github.com/nicolas-gervais/rotten-tomatoes-dataset)

**P.T.O**

4. Textual Data Preprocessing

I will be pre-processing the Stanford Sentiment Treebank:

