# Project Summary:

The aim of this Natural Language Processing (NLP) project is to conduct sentiment analysis on movie reviews. The data preprocessing phase focuses on implementing spell checking techniques to enhance sentiment analysis accuracy. Challenges encountered where proper nouns such as director names, actors, characters, and other entities found in movie reviews were also getting corrected by spell check.

# Spell Checking Approaches and Challenges:

1. **Initial Approach - Pyspellchecker**

* **Approach**: Utilizing the ‘pyspellchecker’ library to correct spelling errors.
* **Challenge**: This method wrongly tries to correct named entities (e.g., actor and director names) to other English words, impacting the context.

**2. Enhanced Approach - Spacy NER Integration**

* **Approach**: Using Spacy's Named Entity Recognition (NER) to identify and ignore recognized entities during spell check.
* **Challenge**: Despite integrating NER, certain character names, production house names, and other entities are still mistakenly corrected.

**3. Ongoing Investigation - Tailoring Existing NER Model**

* **Approach:** Tailoring the existing NER model using additional data from a credits CSV file containing cast and crew information.
* **Challenge:** The cast and crew information in the CSV file is in JSON format, resulting in JSON decoding errors as some property names are not enclosed withing double quotes, need to resolve this issue for adding custom labels into existing NER model.

# Project Next Steps:

* **Refinement of NER Model:** Address formatting issues in the credits CSV file to incorporate specific entity names like cast and crew members into the NER model for improved entity recognition and exclusion from spell checking.
* **Evaluation and Testing**: Rigorous evaluation and testing of the updated spell checking process to ensure context preservation.