CSCE 5290: Natural Language Processing Project Proposal

Title: Ensuring Genuine Feedback from Movie Reviews Dataset

1. Motivation

The identification of fraudulent reviews in movie datasets addresses a common problem: reliability. Differentiating between real and fake online reviews is becoming increasingly important due to their exponential rise. Reviews are crucial for consumers to make informed choices, but fraudulent reviews decrease their confidence. We protect the integrity of internet platforms by identifying and eliminating fraudulent reviews, guaranteeing that consumers get trustworthy information for the movies they choose.

The project's importance comes from how it will affect consumers' decision-making and corporate ethics. False reviews influence opinions, mislead customers, and harm the reputations of brands. We promote a fair marketplace and rebuild trust in online platforms by eliminating fraudulent reviews. Genuineness becomes crucial since it gives customers accurate details and helps firms keep their audience's trust. Additionally, preventing fraudulent reviews improves the quality and analysis of data. For significant insights and well-informed decision-making, reliable datasets are necessary. We improve the accuracy of sentiment analysis and market research by eliminating fraudulent reviews from movie databases. This project improves the integrity of data-driven processes and protects consumer interests and company reputations, which will ultimately benefit stakeholders in a variety of industries.

2. Significance

The project's importance derives from its major impact on audiences and movie industry firms. We guarantee that consumers obtain accurate and reliable information when choosing which movies to see by detecting false reviews in movie datasets. By improving user happiness and experience, promotes trust in online review sites and helps create a more trustworthy and transparent digital environment.

The project also has consequences for companies that work in the film industry. Fake reviews can severely harm the reputation and box office performance of a movie. Filmmakers and production companies can gain a deeper understanding of authentic audience input by efficiently identifying and eliminating fake reviews. This helps them to make well-informed decisions regarding advertising strategies, production investments, and content development. Ultimately, this improves the sector's capacity to produce top-notch films that connect with viewers and ultimately lead to success and sustainability.

3. Objectives

The objective of our project is to conduct an extensive evaluation of various sentiment detection algorithms using the polarity dataset v2.0 to analyze movie reviews and Clearly distinguish between genuine and fake reviews.

Success Criteria:

- To Achieve a similar analysis for at least three sentiment detection algorithms based on the accuracy, precision, recall, and F1 score in identifying fake and genuine movie reviews using the polarity dataset v2.0.
- To enhance the existing sentiment analysis techniques using the polarity dataset v2.0 to Improve scalability.
- To Demonstrate a minimum improvement of 10 percent in the accuracy of sentiment analysis models when compared to the baseline performance metrics through cross-validation on the polarity dataset v2.0.
- To develop a robust and efficient system that is capable of detecting fake reviews in real-time with the insights gained from the polarity dataset v2.0.
- To Implement a functional prototype that achieves real-time processing of at least ten thousand reviews per minute with an accuracy rate exceeding 90% with the validation through rigorous testing on a separate dataset and user feedback.

4. Features

Our project will leverage advanced NLP techniques to analyze the sentiments expressed in movie reviews. By employing sentiment classification strategies, including Support Vector Machines (SVM) and cutting-edge transformer-based models like BERT (Bidirectional Encoder Representations from Transformers), we aim to accurately discern the underlying sentiments in textual data. The initial step involves consolidating the existing segmented data into a comprehensive CSV file to streamline processing. To enhance the precision of our sentiment analysis, we will implement various methods such as tokenization, the use of sentiment lexicons, and sophisticated word embedding techniques, including Word2Vec and GloVe.

Deliverables and Milestones:

- **Sentiment Analysis Model:** The core deliverable is a robust sentiment analysis model. We will train two models, BERT and SVM, on the designated dataset and compare their performance to determine the most effective approach for our application.
- **User Interface:** A key deliverable is the development of a user-friendly interface. This platform will enable users to submit movie reviews and promptly receive sentiment predictions, classifying the reviews as either positive or negative.
- **Comprehensive Report:** A detailed report will be submitted, documenting the project's methodology, findings, and conclusions. This report will provide insights into public reception, box office trends, and critical reception, offering valuable feedback to the film industry.

Project Milestones:

- **Data Acquisition:** Secure the movie review datasets from the designated sources.
- **Data Preprocessing:** Implement preprocessing techniques to clean the data, including tokenization, and removal of stopwords and punctuation, preparing it for analysis.
- **Model Implementation:** Deploy and evaluate both the SVM and BERT models for sentiment analysis.
- **Interface Development:** Construct a user-friendly interface for easy sentiment prediction.
- **Testing and Optimization:** Perform rigorous testing and optimization to enhance model accuracy.
- **Documentation and Submission:** Finalize the project report and documentation for submission.

Distinctive Elements:

- **Our project uniquely compares two models:** BERT, a state-of-the-art deep learning method, and SVM, a conventional machine learning technique. This comparative analysis will enable us to evaluate the strengths and weaknesses of each approach thoroughly. By conducting this comparative study, we aim to identify the most effective sentiment analysis model for movie reviews. This critical evaluation will significantly contribute to the project's success by ensuring the selection of an accurate and efficient model for sentiment analysis.

5. Dataset

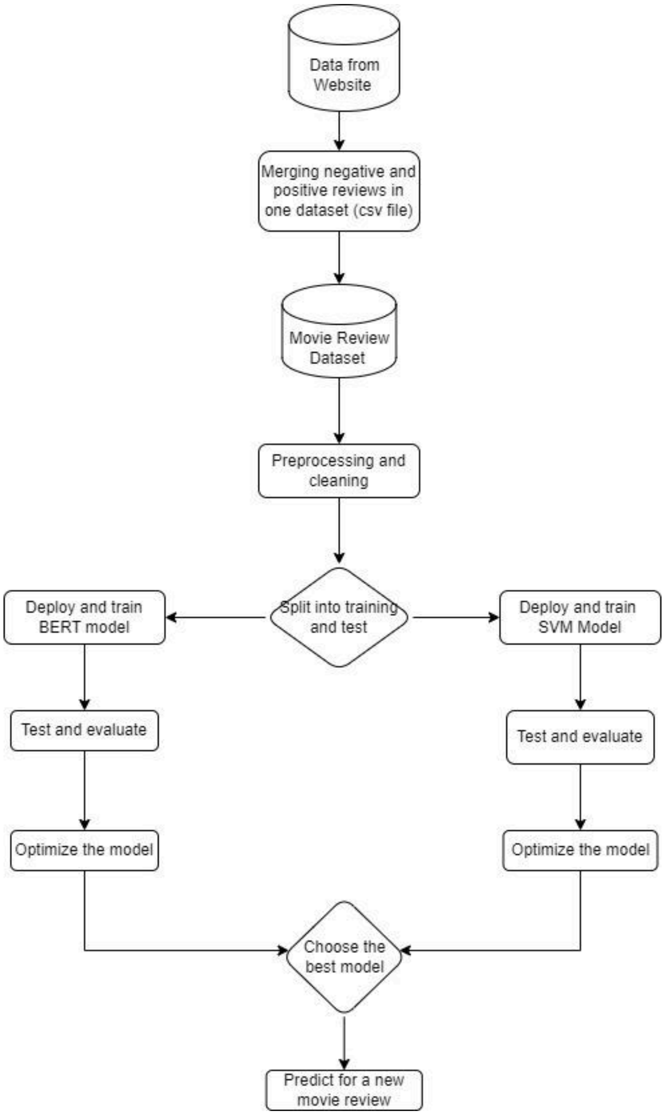
Dataset Details:

We have chosen the movie review dataset provided by Cornell University's Natural Language Processing Group, specifically the polarity dataset v1.1. This dataset, comprising approximately 700 positive and 700 negative processed movie reviews, is meticulously labeled to indicate sentiment polarity. Originally curated by Nathan Treloar and released in November 2002, this dataset underwent careful modifications to enhance its utility, including the removal of non-English or incomplete reviews and adjustments to certain polarity labels to ensure accuracy. Stored in text format, each review is directly associated with its sentiment label, facilitating straightforward analysis.

Preprocessing Steps:

Our preprocessing regimen will involve comprehensive text cleaning to eliminate punctuation and stopwords, coupled with tokenization to dissect the text into analyzable components. Additionally, we may employ lemmatization or stemming techniques to normalize the text further, ensuring the data is primed for efficient and effective sentiment analysis.

6. Visualization


Data from Website
Merging negative and positive reviews in one dataset (csv file)
Movie Review Dataset
Preprocessing and cleaning
Split into training and test
Deploy and train BERT model
Deploy and train SVM Model
Test and evaluate
Test and evaluate
Optimize the model
Optimize the model
Choose the best model
Predict for a new movie review

The flowchart starts from data collection and culminates in the deployment of a refined sentiment analysis model capable of classifying the sentiment of new movie reviews.

Data from Website: This is the initial stage where the raw movie review data is collected from the Cornell University dataset website. It involves downloading the dataset that comprises positive and negative movie reviews, ensuring the data is in a usable format for the subsequent steps.

Merging Negative and Positive Reviews into One Dataset: In this step, separate sets of positive and negative reviews are amalgamated into a singular CSV file. This combined file will serve as the consolidated dataset that facilitates easier access and manipulation during preprocessing and analysis.

Movie Review Dataset: Now that the data is merged, we have a coherent movie review dataset that includes both sentiment polarities. This dataset will be the foundation upon which the preprocessing, model training, and sentiment analysis will be conducted.

Preprocessing and Cleaning: Before feeding the dataset into the models, it must be preprocessed and cleaned. This involves removing any noise that could affect the model's performance, such as punctuation, HTML tags, and stop words. Tokenization, lemmatization, or stemming may also be applied to standardize the text.

Split into Training and Test: The cleaned dataset is then divided into training and test sets. The training set is used to teach the models to recognize sentiment patterns, while the test set is used to evaluate the models' performance on unseen data.

Deploy and Train Models: Two separate paths diverge at this stage:
BERT Model: A sophisticated deep learning model is deployed and trained using the training set. BERT's transformer architecture is designed to capture the context of words in a sentence, making it highly suitable for understanding the nuances of sentiment in text.
SVM Model: A traditional machine learning model is deployed and trained simultaneously. SVMs are effective for classification tasks and provide a baseline against which the BERT model's performance can be compared.
Test and Evaluate: Each model's performance is tested using the test set. Metrics such as accuracy, precision, recall, and F1-score are calculated to evaluate how well each model can classify the sentiment of the reviews.

Optimize the Model: Based on the evaluation metrics, further optimizations are performed on both models to improve their performance. This could involve hyperparameter tuning, using different preprocessing strategies, or enriching the models' training data.

Choose the Best Model: After optimization, the BERT and SVM models are compared to determine which performs better in terms of accuracy and other metrics. The best-performing model is selected for deployment in the sentiment prediction application.

Predict for a New Movie Review: The selected model is then used to make predictions on new movie reviews. Users can input a review into the system, and the model will predict whether the sentiment of the review is positive or negative.