Tammy Chen
301472839
Ling 450

Mid-term Report

Project Title: Sentiment Analysis of Movie Reviews

Team Member(s): Tammy Chen

Objective:

The primary objective of this project is to develop a classifier that automatically categorizes movie reviews into positive, negative, or neutral sentiments. This classifier aims to help viewers quickly assess whether a movie is worth watching and understand audience preferences across different movie genres.

Tasks and Deliverables

Completed Tasks

- Data Collection - Successfully scraped 20 movie reviews from Rotten Tomatoes about Snow White using Python and BeautifulSoup.
- Data Organization and Labeling - Organized collected data into a structured pandas DataFrame AND manually labeled reviews as positive, negative, or neutral.

In Progress Tasks

- Text Processing and Feature extraction - begin exploring linguistic features with spaCy, focusing on adjectives and sentiment-bearing phrases.
- Familiarizing reviews with Nltk sentiment lexicons for tokenization, stopwords and so on.
- Implemented automatic sentiment analysis using the NLTK VADER sentiment analyzer.
- Assessed using metrics like scores of accuracy, precision, recall and F1.
- Assessed labeling reliability using Cohen's Kappa.
- Depending on the data of metrics, analyze movie reviews to judge if a movie is worth watching.

Updated Schedule

Week 8-10: Selected movie, data collection (Completed).

Week 10-12: Feature extraction using spaCy (In progress).

Week 12-13: Preprocessing text using Nltk to tokenize the text, remove step words and Lemmatize the tokens. Develop sentiment classifier using Nltk and machine learning techniques. Calculating metrics like accuracy, precision, recall, F1 and Cohen's Kappa. Analyze movie sentiments across general information and data (In progress).

Tammy Chen
301472839
Ling 450

Week 14: Discover and improve the content and vulnerabilities of the project. Try to match the content and output as much as possible. Prepare a final report (Planning).

New Tasks or Adjustments:

• No new tasks or deliverables have been added since the initial proposal.

• The division of labor remains unchanged, with all tasks completed individually.

Risks and Challenges (Currently)

• Understand the structure of the Rotten Tomatoes website. Find the specific tags that identify each review on the page in the large block of HTML tags on the website.

• Anticipating challenges in accurately identifying positive, negative, and neutral comments.

Communication Plan

• Updates and progress tracking maintained via Github, Jupyter notebook and google docs.