Tammy Chen
301472839
Ling 450

Ling 450 Final Project 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.

Description of Project:

This project aims to develop a text classifier that identifies positive, negative or neutral sentiments in film reviews. Reviews will be gathered through web scraping from popular film review websites, ensuring a balanced dataset covering various film genres. I will utilize BeautifulSoup for web scraping, Nltk for natural language processing tasks such as text tokenization, removing stop words, lemmatizing the token and sentiment lexicons for advanced text preprocessing and feature extraction. Model accuracy will be assessed using metrics like accuracy, precision, recall and F1 score. Manually labeled reviews and assessed labeling reliability using Cohen's Kappa. Analyze movie reviews to judge if a movie is worth watching.

Tasks and Deliverables

- 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.
- Text Processing and Feature extraction - begin exploring linguistic features with spaCy, focusing on adjectives and sentiment-bearing phrases(Removed).
- Familiarizing reviews with Nltk sentiment lexicons for tokenization, stopwords and so on.
- Implemented automatic sentiment analysis using the NLTK VADER compound 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.

Final Timeline Of Schedule

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

Week 10-12: Feature extraction using spaCy (Revised, removed).

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 (Completed).

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 (Completed).

Results and Discussion

Adjusting the threshold methods of sentiment analysis from only labelling positive and negative to labelling positive, negative, or neutral, which the change improved classification accuracy from 45% to 65%. Assessed labeling reliability using Cohen's Kappa, resulting in a high consistency score of 0.91. In finding, my sentiment analyzer showed 11 positive, 8 negative, and 1 neutral review for Disney's Snow White (2025), indicating mixed but slightly positive sentiment. This suggests the movie may be worth watching, particularly for audiences interested in strong performances and modern adaptations of classics. The project helped me review the learning of web scraping and gained practical experience with sentiment analysis and classifier evaluation. Refining classification tools and methodologies to achieve reliable results. For the future improvement of the sentiment analyzer, I want to expand data collection to include more reviews. Using spaCy to identify complex sentiments like sarcasm or subtle negative phrasing and extract additional linguistic features.

Resources:

- Python Libraries included Beautifulsoup, pandas DataFrame and Nltk
- Data of movie reviews from Rotten Tomatoes website
- Jupyter Notebook, Github and Google Docs for documentation
- Ling 250&450 materials and tutorials from Prof.Taboada in SFU

Tammy Chen
301472839
Ling 450

• Online NLTK Sentiment Analysis Tutorial for Beginners

Risks and Challenges (Currently)

• The review data scraped from the website is real but may fluctuate according to the actual situation. Since the website content is constantly changing, it also poses difficulties for the act of seeking data evidence.

• Uncertainty regarding the accuracy of the obtained data.

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.

Communication Plan

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

Tammy Chen
301472839
Ling 450

References

DataCamp. (n.d.). *Text analytics for beginners using NLTK*. https://www.datacamp.com/tutorial/text-analytics-beginners-nltk

Geetha, R. S. (n.d.). *VADER: A comprehensive guide to sentiment analysis in Python*. Medium. https://medium.com/@rslavanyageetha/vader-a-comprehensive-guide-to-sentiment-analysis-in-python-c4f1868b0d2e

Hutto, C. J. (n.d.). *vaderSentiment*. GitHub. https://github.com/cjhutto/vaderSentiment

Laxmimerit. (n.d.). *NLP Tutorial 8 – Sentiment classification using SpaCy for IMDB and Amazon review dataset*. GitHub. https://github.com/laxmimerit/NLP-Tutorial-8---Sentiment-Classification-using-SpaCy-for-IMDB-and-Amazon-Review-Dataset

Rotten Tomatoes. (n.d.). *Disney's Snow White - Reviews*. Rotten Tomatoes. https://www.rottentomatoes.com/m/disneys_snow_white/reviews

Taboada, M. (n.d.). *LING 450*. GitHub. https://github.com/maitetaboada/Ling450