CPE 593 - Milestone 1

Group Members:
Jude Eschete
Ella Disanti
Raymond Donkemezuo

Github Link: https://github.com/JEschete2651/finalProject

I. Introduction

The objective of this project is to perform sentiment analysis on movie reviews using a lexicon-based approach, as detailed in the project topic description. This report aims to provide a clear roadmap for completing the remaining tasks required for Milestone 1, which are the assignment of workload and tasks and the creation of a comprehensive project plan. By outlining these elements, the project team will be well-prepared to move forward with the implementation and analysis phases.

II. Detailed Breakdown of Remaining Milestone Tasks

A. Workload and Task Assignment

For the successful execution of a lexicon-based sentiment analysis project, several key roles and responsibilities are essential. These roles will ensure that all aspects of the project are covered efficiently. Given the team members Jude Eschete, Ella Desanti, and Raymond Donkemezuo, and considering their equal skill levels, the roles are assigned as follows to ensure a balanced workload:

- **Jude Eschete:** Overall project management, ensuring adherence to timelines, facilitating communication, guiding the project, and evaluating the outcomes. Jude will be responsible for the final deliverable of Milestone 1, the progress report, and for analyzing the distribution of sentiment scores and assessing performance later in the project.[1]
- **Ella Desanti:** Focus on the core technical aspects. Ella will investigate, compare, and select the most suitable sentiment lexicon(s) and implement the logic for sentiment scoring and aggregation.[1] Examples of lexicons to be researched include VADER, SentiWordNet, Bing, and NRC.[2]

- **Raymond Donkemezuo:** Handle data preparation and record-keeping. Raymond will focus on preparing the movie review data for analysis (tokenization, cleaning, etc.) and maintaining thorough documentation throughout the project, including the project plan

and progress reports.[1]

To achieve the goals of Milestone 1 and set the stage for the rest of the project, the following specific tasks need to be accomplished, assigned as follows:

- **Milestone 1 Specific Tasks:**
  - **Lexicon Research:** Ella Desanti will conduct a thorough investigation into various sentiment lexicons (VADER, SentiWordNet, Bing, NRC), evaluating their characteristics and suitability for movie reviews.[2]
  - **Data Preprocessing Plan:** Raymond Donkemezuo will outline the specific steps required to prepare movie review text (tokenization, punctuation handling, lowercasing).[1]
  - **Sentiment Aggregation Plan:** Ella Desanti will develop a basic plan for how sentiment scores will be aggregated (e.g., summing word scores) to determine the overall review sentiment.[1]
  - **Project Plan Structure:** Jude Eschete and Raymond Donkemezuo will collaborate to create the structure for the project plan document, outlining phases, tasks, and responsibilities.
  - **Final Task Assignment & Report Compilation:** Jude Eschete will finalize the assignment details, define expected outputs, set deadlines, and compile the Progress Report v1 for Milestone 1.

- **Tasks Beyond Milestone 1 (for overall project planning):**
  - **Implementation (Sentiment Analysis):** Ella Desanti will implement the chosen lexicon-based approach (e.g., in Python), potentially referencing existing examples.[6]
  - **Dataset Acquisition:** Raymond Donkemezuo and Ella Desanti will collaborate to identify and obtain a suitable movie review dataset (e.g., from IMDb, Rotten Tomatoes, Amazon).[7]
  - **Data Preprocessing (Execution):** Raymond Donkemezuo will process the chosen dataset using the steps outlined in Milestone 1.
  - **Sentiment Scoring (Execution):** Ella Desanti will apply the selected lexicon and aggregation method to calculate sentiment scores for the dataset.
  - **Negation Handling:** Ella Desanti will explore and plan basic strategies for handling negation in the text.[5]
  - **Results Evaluation:** Jude Eschete will evaluate the results, potentially comparing them against a small manually labeled subset.
  - **Documentation:** Raymond Donkemezuo will ensure all processes, findings, and challenges are thoroughly documented.

Effective task division is crucial for project efficiency. Jude Eschete, should ensure that all assigned tasks are specific, measurable, achievable, relevant, and time-bound. The initial task assignment for Milestone 1 prioritizes research and planning activities. This foundational work will be instrumental in guiding the subsequent implementation phase. The role of Ella Desanti as Lexicon Researcher is particularly important at this stage. The success of a lexicon-based sentiment analysis approach is intrinsically linked to the quality and relevance of the chosen sentiment lexicon.[3] Therefore, dedicating effort to thoroughly investigate and select the most appropriate lexicon ensures that the project's foundation is robust, upon which the subsequent tasks of data processing and sentiment scoring will be built.

B. Project Planning

A well-defined project plan is essential for organizing and executing the lexicon-based sentiment analysis project effectively. The project can be broadly divided into the following phases, which align with typical natural language processing project workflows [1]:

- **Phase 1: Project Setup and Planning (Current Milestone)**
  - Team formation has been completed (Jude Eschete, Ella Desanti, Raymond Donkemezuo).
  - The project topic has been selected.
  - Workload and task assignment are defined in this document.
  - Project planning is the focus of this part of the milestone.
  - The GitHub repository has been successfully set up.

- **Phase 2: Data Acquisition and Preprocessing**
  - This phase will involve identifying and obtaining a suitable movie review dataset from the various publicly available sources.[1]
  - Once the dataset is acquired, data cleaning steps will be implemented to handle any missing data, duplicate entries, or inconsistencies in the data.[1]
  - The text data will then undergo preprocessing, which includes tokenization (splitting text into words), converting text to lowercase, removing punctuation and special characters, filtering out stop words (common words that may not carry sentiment), and applying stemming or lemmatization to reduce words to their base form.[1] Common preprocessing techniques include tokenization, normalization, and stop word removal.[27]

- **Phase 3: Lexicon Application and Sentiment Scoring**
    - In this phase, the chosen sentiment lexicon will be loaded into the project's software environment.
    - The project will then iterate through each of the preprocessed movie reviews in the dataset.
    - For every word in each review, the system will look up its corresponding sentiment score in the loaded lexicon.[2]
    - The sentiment scores for all the words in a review will be aggregated using a selected method, such as summation or averaging, to determine the overall sentiment score for that review.[1]

- **Phase 4: Results Analysis and Evaluation**
    - The distribution of the calculated sentiment scores will be analyzed to understand the overall sentiment expressed in the movie reviews (positive, negative, or neutral).
    - Optionally, for an initial evaluation, the results can be compared with a small subset of movie reviews that have been manually labeled with sentiment or with the known ratings of the movies, if available, to get a preliminary sense of the accuracy of the lexicon-based approach.[41] Evaluation metrics such as accuracy, precision, recall, and F1-score can be used.[41]
    - Any limitations or areas for potential improvement in the lexicon-based method will be identified.

- **Phase 5: Documentation and Reporting**
    - Throughout the project, all steps taken, decisions made, and results obtained will be meticulously documented.
    - The progress report for Milestone 1 will be prepared, detailing the status of the project up to this point.
    - A final comprehensive report will be created at the conclusion of the project, summarizing the entire process, findings, and conclusions.

For the immediate future, leading up to the next milestone, the following sequence of tasks is proposed:

1. Finalize the workload and task assignments for Milestone 1, ensuring each team member (Jude, Ella, Raymond) has clear responsibilities.
2. Ella Desanti (Lexicon Researcher) should conduct in-depth research on the available sentiment lexicons, carefully evaluating their features and selecting the most appropriate one(s) for the project's focus on movie reviews.
3. Raymond Donkemezuo (Data Preprocessor) should outline the detailed steps involved in preprocessing the movie review data, including the specific techniques for tokenization, handling punctuation, lowercasing, stop word removal, and stemming/lemmatization.
4. Ella Desanti (Sentiment Analyzer/Scorer) needs to develop a clear plan for how the sentiment scores obtained from the lexicon will be aggregated to produce an overall sentiment score for each movie review.
5. All the above information, including the task assignments, lexicon selection rationale, preprocessing plan, and aggregation strategy, should be documented by Raymond Donkemezuo (Documentation Lead) and compiled by Jude Eschete (Project Lead) into the Progress Report v1, which is a deliverable for Milestone 1.
6. The GitHub repository should be organized by Raymond Donkemezuo (Documentation Lead) with a clear directory structure to accommodate the next phases of the project. This might include directories for data, code, and documentation.

III. Key Considerations for Lexicon-Based Sentiment Analysis

A. Lexicon Selection

The selection of an appropriate sentiment lexicon is a crucial decision that will significantly influence the outcome of the sentiment analysis.[3] Several readily available lexicons can be considered for this project:

- **VADER (Valence Aware Dictionary and sEntiment Reasoner):** This lexicon is specifically designed for sentiment analysis in social media texts and is known for its ability to handle slang, emoticons, and sentiment intensity.[3] While the project focuses on movie reviews, VADER's sensitivity to nuanced language might be beneficial.
- **SentiWordNet:** As mentioned in the project description, SentiWordNet is a lexical resource that assigns scores for positivity, negativity, and objectivity to words (synsets) in WordNet.[3] It offers a more granular view of sentiment and is available through the Natural Language Toolkit (NLTK) in Python.[45]
- **Bing lexicon:** This lexicon provides a simple classification of words as either positive or negative.[2] It is straightforward to use and can be a good starting point for basic polarity detection.
- **NRC Emotion Lexicon:** This lexicon goes beyond simple polarity and categorizes words based on both positive and negative sentiments, as well as eight basic emotions like anger, fear, joy, and sadness.[2] The NRCLex lexicon, which is based on this, is available for Python.[46]
- **AFINN:** This lexicon provides a list of English words, each rated with a valence score ranging from -5 (very negative) to +5 (very positive), indicating the intensity of sentiment.[33]
- **MPQA Opinion Corpus:** This corpus includes lists of positive and negative words and can be used as a sentiment lexicon.[5]

When making a decision about which lexicon to use, several factors should be taken into account:

- **Domain Specificity:** While some general-purpose lexicons exist, a lexicon that is specifically tailored to the domain of movie reviews might yield more accurate results. However, general lexicons can still be effective.[48]
- **Sentiment Categories:** The project aims to classify sentiment as positive, negative, or neutral. The chosen lexicon should ideally support these categories, either directly or through its scoring system.
- **Scoring Range:** Different lexicons use different scoring ranges. Understanding this range is important for interpreting the aggregated sentiment scores.
- **Handling of Nuances:** Some lexicons are better at capturing the intensity of sentiment than others.[45] This could be important for distinguishing between reviews that are mildly positive versus overwhelmingly positive.
- **Size and Coverage:** The lexicon should have a broad enough vocabulary to cover most of the words likely to appear in movie reviews. A limited vocabulary size can be a

drawback.[32]

- **Language:** Given that the movie reviews are expected to be in English, the chosen lexicon must be appropriate for the English language.[3]

The selection of the lexicon is a critical factor that will significantly impact the results of the sentiment analysis. Therefore, it is essential to carefully evaluate the available options based on the specific characteristics of movie review language and the project's goals. Different lexicons are constructed with varying purposes and methodologies.[3] For example, VADER is optimized for social media, while SentiWordNet offers a more comprehensive semantic resource. Applying a lexicon that is not well-suited to the movie review domain could lead to less accurate sentiment classifications.[48] Thus, a thorough comparison of the features of different lexicons will provide a structured overview, allowing the team to weigh the pros and cons of each option based on the required sentiment categories, scoring mechanisms, and relevance to the movie review domain.

Lexicon Name	**Source**	**Sentiment Categories**	**Scoring Range**	**Domain Focus**	**Example Words and Scores (Illustrative)**
VADER	Python library	Positive, Negative, Neutral	-4 to +4 (compound score)	Social Media	great: +3.1, terrible: -3.2
SentiWordNet	NLTK	Positive, Negative, Objective	0 to 1 for each	General	good: pos=0.875, neg=0.125, obj=0.0
Bing	R package (tidytext)	Positive, Negative	Implicit (+1/-1)	General	happy: positive, sad: negative
NRC Emotion	R package (tidytext), NRCLex (Python)	Positive, Negative, 8 Emotions	Implicit (+1/-1)	General	happy: positive, joy; angry: negative, anger
AFINN	R package (tidytext)	Polarity	-5 to +5	General	love: 3, hate: -3
MPQA Opinion	Website	Positive, Negative	Implicit	General	good, bad

B. Data Processing

Properly processing the movie review data is a fundamental step that ensures the lexicon-based sentiment analysis is as accurate and effective as possible. The necessary steps generally include:

- **Tokenization:** This involves breaking down each movie review from a continuous string of text into a sequence of individual words, or tokens.[1] This is the first step in making the text amenable to analysis at the word level.
- **Lowercasing:** Converting all the text in the reviews to lowercase ensures consistency. This way, the word "Good" and "good" will be treated as the same word when looking up sentiment in the lexicon.[1]
- **Removal of Punctuation and Special Characters:** Punctuation marks and special characters generally do not carry sentiment in themselves and can be removed to reduce noise in the data and focus on the actual words.[1]
- **Stop Word Removal:** Stop words are common words in a language, such as "the," "a," "is," "and," which often do not contribute significantly to the sentiment of a text. Removing these words can help focus the analysis on more sentiment-bearing terms.[1] However, it is important to be cautious as some stop words can be part of sentiment-carrying phrases, and their removal might unintentionally alter the meaning.[44]
- **Stemming or Lemmatization:** These are techniques used to reduce words to their base or root form. Stemming typically involves removing suffixes, while lemmatization aims to find the dictionary form (lemma) of a word.[1] For example, "running," "runs," and "ran" might all be reduced to "run." This helps in grouping different forms of the same word together, but it's worth noting that this might affect the ability to directly match words to some lexicons that rely on specific word forms.[44]

Ensuring that the movie review data is clean and well-structured through these preprocessing steps is crucial for the accuracy of the subsequent sentiment analysis.[1] By removing noise and standardizing the text, the chances of accurately matching words with their sentiment scores in the chosen lexicon are significantly improved.

Preprocessing Step	Description	Example (Before)	Example (After)	Rationale
Tokenization	Breaking text into individual words or tokens.	This movie was fantastic!	\	Prepares the text for word-level analysis.
Lowercasing	Converting all text to lowercase.	This Movie Was Fantastic!	this movie was fantastic!	Ensures consistency in word matching, regardless of capitalization.
Removal of Punctuation	Eliminating punctuation marks and special characters.	This movie was fantastic!	This movie was fantastic	Punctuation usually does not carry sentiment in a lexicon-based approach.
Stop Word Removal	Filtering out common, non-sentiment-bearing words.	This movie was fantastic!	movie fantastic	Focuses analysis on words that are more likely to express sentiment.
Lemmatization (or Stemming)	Reducing words to their base or root form.	This movie was running fantastic	this movie was run fantastic	Groups different inflections of the same word together, improving matching with the lexicon.

C. Sentiment Score Aggregation Methods

Once the sentiment scores for individual words in a movie review are retrieved from the chosen lexicon, these scores need to be combined to determine the overall sentiment of the review. Several methods can be used for this aggregation [3]:

- **Summation:** This is a straightforward method where the sentiment scores of all the words in a review that are present in the lexicon are added together. The total sum represents the overall sentiment score of the review.[1]
- **Averaging:** In this method, the sentiment scores of the lexicon words in a review are averaged. This can help to normalize the score by the number of sentiment-bearing words and might be less influenced by the length of the review.[3]
- **Weighted Summation:** This approach involves assigning different weights to words based on factors such as their intensity of sentiment or their position in the sentence. For example, intensifiers like "very" might increase the weight of the following sentiment word.[52]
- **Considering Frequency:** The frequency with which a sentiment-bearing word appears in a review can also be taken into account. A word that appears multiple times might have a greater impact on the overall sentiment.[5] However, repeated words could also skew the results.[2]
- **Normalization:** After aggregating the scores, the final sentiment score can be normalized to a specific range, such as -1 to +1 or 0 to 1, to facilitate comparison across different reviews.[37]

The choice of aggregation method can significantly influence the final sentiment score assigned to a movie review.[54] Therefore, it might be beneficial to experiment with different methods to see which one yields the most meaningful and accurate results for the project.

D. Handling Negation

Handling negation is a critical aspect of sentiment analysis, as the presence of negative words can reverse the sentiment of associated words.[24] Several strategies can be employed to address negation in a lexicon-based approach [2]:

- **Identifying Negation Words:** The first step is to create or use a predefined list of common negation words in the English language, such as "not," "no," "never," "neither," and contractions like "n't."
- **Inverting Polarity:** A common technique is to invert the polarity of a sentiment-bearing word if it is preceded by a negation word within a certain proximity, often referred to as a window.[20] For example, if the word "good" (positive) appears after "not," its sentiment score could be flipped to negative. The size of this window (e.g., the next one or two words) needs to be determined.
- **Handling Negative Phrases:** Some approaches involve recognizing common negative phrases as single units with a negative sentiment. For instance, "not good" can be treated as a single negative expression.[20]
- **Considering the Scope of Negation:** More advanced techniques attempt to determine the scope of influence of a negation word within a sentence. This involves identifying which words in the sentence are affected by the negation.[20] Linguistic analysis, such as part-of-speech tagging and dependency parsing, can be helpful in this regard.
- **Using N-grams:** Considering sequences of words (n-grams) can help capture the sentiment of negated phrases. For example, the bigram "not good" can be directly associated with a negative sentiment, rather than analyzing "not" and "good" separately.[24]

Accurately handling negation is a complex task in natural language processing.[2] However, implementing even basic negation handling strategies can significantly improve the accuracy of lexicon-based sentiment analysis by preventing misclassification of sentiment in the presence of negative words.

E. Potential Challenges and Solutions

While lexicon-based sentiment analysis offers a foundational approach, it is important to be aware of its inherent limitations [1]:

- **Context-Dependent Sentiment:** The meaning and sentiment of a word can vary significantly depending on the context in which it is used. Lexicon-based methods, which primarily look at individual words, may struggle to capture these contextual nuances.[1]
- **Sarcasm and Irony:** When sentiment is expressed through sarcasm or irony, the literal meaning of the words often contradicts the intended sentiment. Lexicon-based approaches typically find it difficult to detect such subtleties.[1]
- **Negation:** As previously discussed, accurately handling negation to understand its impact on sentiment can be challenging.[2]
- **Domain Specificity:** Sentiment lexicons are often developed for specific domains. A lexicon trained on general text or social media might not perform optimally when applied to movie reviews, which may have their own specific vocabulary and sentiment expressions.[48]
- **Evolving Language:** Languages are constantly evolving, with new words, slang terms, and abbreviations emerging. Existing lexicons might not contain these newer terms, leading to inaccuracies in sentiment analysis.[32]
- **Subjectivity vs. Objectivity:** Distinguishing between subjective statements that express opinions and objective statements that present facts can be difficult for lexicon-based methods.[28]
- **Intensity of Sentiment:** Simply classifying sentiment as positive or negative might not capture the intensity or degree of the emotion expressed.[45]
- **Cultural and Linguistic Differences:** Sentiment can be expressed differently across various languages and cultures. A lexicon developed for one language might not be directly applicable to another.[3]
- **Sentiment Bias:** Lexicon-based systems might exhibit a tendency to produce strong positive or negative sentiment scores, potentially overemphasizing sentiment where a more neutral classification might be appropriate.[69]

To mitigate these challenges and improve the accuracy of lexicon-based sentiment analysis, several solutions can be considered [3]:

- **Utilizing Domain-Specific Lexicons:** If available, using a sentiment lexicon that has been specifically created or adapted for the domain of movie reviews could enhance accuracy.
- **Combining with Other Techniques:** Exploring hybrid approaches that integrate lexicon-based methods with machine learning techniques can leverage the strengths of both. Machine learning models can help in understanding context and nuances that lexicon-based methods might miss.[3]
- **Manual Refinement of Lexicon:** If time allows, manually reviewing and refining the chosen lexicon by adding movie-specific terms or adjusting sentiment scores can improve its performance for this particular domain.
- **Implementing Advanced Negation Handling:** Employing more sophisticated techniques to accurately identify the scope and impact of negation in movie reviews.
- **Incorporating Contextual Information:** Exploring methods to consider the context of words, such as analyzing n-grams (sequences of words) or using part-of-speech tags to better understand how words are used in sentences.[3]
- **Employing Ensemble Methods:** Combining the results from multiple different lexicon-based approaches and potentially machine learning methods can lead to more robust and accurate sentiment predictions.[83]

Being aware of these inherent limitations of lexicon-based sentiment analysis is crucial for setting realistic expectations for the project's accuracy. It will also help in critically interpreting the results and in considering more advanced techniques if the initial lexicon-based approach does not meet the desired level of performance.

IV. Movie Review Datasets

For this project, several publicly available movie review datasets can be utilized [1]:

- **IMDb Movie Reviews Dataset:** This dataset contains a substantial number of movie reviews (50,000) that are labeled as either positive or negative. It is a popular choice for sentiment analysis tasks and is readily available on platforms like Kaggle.[9]
- **Rotten Tomatoes Movie Reviews Dataset:** This corpus provides movie reviews with sentiment labels on a more granular scale, ranging from negative to positive across five values. It is also available on Kaggle and other online repositories.[8]
- **Movie Review Data (Cornell University):** Researchers at Cornell University have compiled several collections of movie review documents that are labeled with respect to their overall sentiment polarity (positive or negative) or subjective rating. These datasets are frequently used in sentiment analysis research.[11]
- **MovieLens Datasets:** While primarily known for movie ratings, some of the MovieLens datasets might include textual reviews or user-generated tags that could potentially be used for sentiment analysis, although the focus is typically on collaborative filtering and recommendation systems.[14]
- **Amazon Movie Reviews:** This is a very large dataset containing millions of movie reviews from Amazon customers. It includes product and user information, ratings, and the review text itself.[18]
- **1000 Movie Reviews (Mendeley Data):** This is a smaller, manually created dataset containing 1000 movie reviews along with attached ratings and sentiment polarity labels, which could be useful for initial testing or evaluation.[13]

The availability of these diverse datasets, varying in size and labeling granularity, allows for flexibility in choosing one that best suits the project's specific needs and the team's computational resources. The presence of labeled data, even if the primary approach is lexicon-based, can be valuable for preliminary evaluation of the system's performance or for potential future exploration of supervised learning methods.

Works cited

1. Sentiment Analysis in NLP: Key Techniques and Insights - Sapien, accessed April 8, 2025, https://www.sapien.io/blog/sentiment-analysis-in-nlp
2. Lexicon-Based Sentiment Analysis Using R | by Okan Bulut | TDS Archive - Medium, accessed April 8, 2025, https://medium.com/data-science/lexicon-based-sentiment-analysis-using-r-5c1db85984a1
3. Simple Guide On How To Do Lexicon Based Sentiment Analysis For Your Business, accessed April 8, 2025, https://numerous.ai/blog/lexicon-based-sentiment-analysis
4. Lexicon-Based Sentiment Analysis Using R - Okan Bulut, accessed April 8, 2025, https://okan.cloud/posts/2024-02-09-lexicon-based-sentiment-analysis-using-r/
5. Lexicon-based sentiment analysis: What it is & how to conduct one - KNIME, accessed April 8, 2025, https://www.knime.com/blog/lexicon-based-sentiment-analysis
6. Srushti25/Sentiment-Analysis-Lexicon-approach ... - GitHub, accessed April 8, 2025, https://github.com/Srushti25/Sentiment-Analysis-Lexicon-approach
7. 10 Sentiment Analysis Project Ideas with Source Code [2025] - ProjectPro, accessed April 8, 2025, https://www.projectpro.io/article/sentiment-analysis-project-ideas-with-source-code/518
8. Sentiment Analysis on Movie Reviews - Kaggle, accessed April 8, 2025, https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews
9. IMDB Dataset of 50K Movie Reviews - Kaggle, accessed April 8, 2025, https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
10. Sentiment Analysis - Stanford AI Lab, accessed April 8, 2025, https://ai.stanford.edu/~amaas/data/sentiment/
11. MR Dataset - Papers With Code, accessed April 8, 2025, https://paperswithcode.com/dataset/mr
12. Movie Review Data - CS@Cornell, accessed April 8, 2025, https://www.cs.cornell.edu/people/pabo/movie-review-data/
13. 1000 Movie Reviews (Review + Attached rating + Sentiment polarity) for Reputation Generation - Mendeley Data, accessed April 8, 2025, https://data.mendeley.com/datasets/38j8b6s2mx/1
14. Where do I get good movie review rating prediction dataset? - Reddit, accessed April 8, 2025, https://www.reddit.com/r/datasets/comments/6rqv49/where_do_i_get_good_movie_review_rating/
15. IMDb Movie Reviews Dataset | IEEE DataPort, accessed April 8, 2025, https://ieee-dataport.org/open-access/imdb-movie-reviews-dataset
16. MovieLens | GroupLens, accessed April 8, 2025, https://grouplens.org/datasets/movielens/
17. The Movies Dataset - Kaggle, accessed April 8, 2025,

https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset
18. SNAP: Web data: Amazon movie reviews - Stanford Network Analysis Project, accessed April 8, 2025, https://snap.stanford.edu/data/web-Movies.html
19. There was an IMDb dataset on kaggle that had detailed ratings breakdown of all movies and was later removed, since then i have not found anything like it. - Reddit, accessed April 8, 2025, https://www.reddit.com/r/datasets/comments/13nqrc8/there_was_an_imdb_dataset_on_kaggle_that_had/
20. personal.eur.nl, accessed April 8, 2025, https://personal.eur.nl/hogenboom/files/smc11-sentineg.pdf
21. When to use negation handling in sentiment analysis?, accessed April 8, 2025, https://analyticsindiamag.com/ai-trends/when-to-use-negation-handling-in-sentiment-analysis/
22. Negation handling for sentiment analysis task: approaches and performance analysis, accessed April 8, 2025, https://www.researchgate.net/publication/379895267_Negation_handling_for_sentiment_analysis_task_approaches_and_performance_analysis
23. Using Meaning Specificity to Aid Negation Handling in Sentiment Analysis, accessed April 8, 2025, https://sites.socsci.uci.edu/~lpearl/CoLaLab/papers/Hii2019_MeaningSpecNegSent.pdf
24. Negation Handling in Machine Learning-Based Sentiment Classification for Colloquial Arabic - arXiv, accessed April 8, 2025, https://arxiv.org/pdf/2107.11597
25. Negation handling in sentiment analysis - python - Stack Overflow, accessed April 8, 2025, https://stackoverflow.com/questions/29374157/negation-handling-in-sentiment-analysis
26. Negation Handling in Sentiment Analysis: Best Practices - Insight7, accessed April 8, 2025, https://insight7.io/negation-handling-in-sentiment-analysis-best-practices/
27. What is Sentiment Analysis? Guide, Tools, Examples | Appinio Blog, accessed April 8, 2025, https://www.appinio.com/en/blog/market-research/sentiment-analysis
28. Sentiment Analysis: A Complete Guide [Updated for 2025], accessed April 8, 2025, https://careerfoundry.com/en/blog/data-analytics/sentiment-analysis/
29. Sentiment Analysis Using Natural Language Processing (NLP) | by Robert De La Cruz, accessed April 8, 2025, https://medium.com/@robdelacruz/sentiment-analysis-using-natural-language-processing-nlp-3c12b77a73ec
30. Preprocessing Steps for Natural Language Processing (NLP): A Beginner's Guide - Medium, accessed April 8, 2025, https://medium.com/@maleeshadesilva21/preprocessing-steps-for-natural-language-processing-nlp-a-beginners-guide-d6d9bf7689c9
31. A complete guide to Sentiment Analysis approaches with AI - Thematic, accessed April 8, 2025, https://getthematic.com/sentiment-analysis
32. How does sentiment analysis work : lexicon-based method | by nabila, accessed

April 8, 2025, https://ai.plainenglish.io/how-does-sentiment-analysis-work-lexicon-based-method-e392c6263e82
33. Lexicon-Based Sentiment Analysis: A Comprehensive Guide - P3MPI, accessed April 8, 2025, https://p3mpi.uma.ac.id/2024/05/10/lexicon-based-sentiment-analysis-a-comprehensive-guide/
34. Lexicon-Based Sentiment Analysis Using R | Towards Data Science, accessed April 8, 2025, https://towardsdatascience.com/lexicon-based-sentiment-analysis-using-r-5c1db85984a1/
35. www.displayr.com, accessed April 8, 2025, https://www.displayr.com/how-to-calculate-sentiment-scores-for-open-ended-responses-in-displayr/#:~:text=Each%20word%20is%20then%20assigned,not%20good%22%20becomes%20negative).
36. How to Calculate Sentiment Scores for Open-Ended Responses in ..., accessed April 8, 2025, https://www.displayr.com/how-to-calculate-sentiment-scores-for-open-ended-responses-in-displayr/
37. Different Methods for Calculating Sentiment of Text - Analytics Vidhya, accessed April 8, 2025, https://www.analyticsvidhya.com/blog/2021/12/different-methods-for-calculating-sentiment-score-of-text/
38. medium.com, accessed April 8, 2025, https://medium.com/illumination/top-5-techniques-for-sentiment-analysis-in-natural-language-processing-c07ba5b83f64#:~:text=To%20perform%20sentiment%20analysis%20using,or%20by%20taking%20the%20average.
39. Top 5 Techniques for Sentiment Analysis in Natural Language Processing - Medium, accessed April 8, 2025, https://medium.com/illumination/top-5-techniques-for-sentiment-analysis-in-natural-language-processing-c07ba5b83f64
40. 2 Sentiment analysis with tidy data - Text Mining with R, accessed April 8, 2025, https://www.tidytextmining.com/sentiment
41. Lexicon and Deep Learning-Based Approaches in Sentiment Analysis on Short Texts, accessed April 8, 2025, https://www.scirp.org/journal/paperinformation?paperid=130385
42. Lexicon and Deep Learning-Based Approaches in Sentiment Analysis on Short Texts - Scientific Research, accessed April 8, 2025, https://www.scirp.org/pdf/jcc_2024010516374080.pdf
43. The Implementation Cycle in Applied Natural Language Processing (NLP) - Deepset, accessed April 8, 2025, https://www.deepset.ai/blog/the-implementation-cycle-in-applied-nlp
44. Sentiment Analysis (Lexicons) - RPubs, accessed April 8, 2025, https://rpubs.com/chelseyhill/676279
45. Sentiment Analysis - The Lexicon Based Approach - AlphaBOLD, accessed April 8,

2025, https://www.alphabold.com/sentiment-analysis-the-lexicon-based-approach/
46. Process of lexicon-based sentiment analysis. | Download Scientific Diagram - ResearchGate, accessed April 8, 2025, https://www.researchgate.net/figure/Process-of-lexicon-based-sentiment-analysis_fig3_367973437
47. Sentiment Analysis of Online Reviews with Different Lexicons using R, accessed April 8, 2025, https://blog.marketingdatascience.ai/sentiment-analysis-of-online-reviews-with-different-lexicons-using-r-bc726649c8ef
48. Lexicon-Based Approach in Sentiment Analysis | MiaRec - Blog, accessed April 8, 2025, https://blog.miarec.com/lexicon-based-vs-ml-based-sentiment-analysis
49. Overcoming Sentiment Analysis Challenges with Machine Learning, accessed April 8, 2025, https://www.cornerstone.com/data-science-center/overcoming-sentiment-analysis-challenges-with-machine-learning/
50. Natural Language Processing Functionality in AI - Turing, accessed April 8, 2025, https://www.turing.com/kb/natural-language-processing-function-in-ai
51. How to Build an NLP Model Step by Step using Python? - ProjectPro, accessed April 8, 2025, https://www.projectpro.io/article/how-to-build-an-nlp-model-step-by-step-using-python/915
52. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons - ResearchGate, accessed April 8, 2025, https://www.researchgate.net/publication/316260588_Lexicon-based_sentiment_analysis_Comparative_evaluation_of_six_sentiment_lexicons
53. Sentiment Score: What It Is and How to Calculate It - AlphaSense, accessed April 8, 2025, https://www.alpha-sense.com/blog/engineering/sentiment-score/
54. (PDF) An improved evidence-based aggregation method for sentiment analysis, accessed April 8, 2025, https://www.researchgate.net/publication/331842248_An_improved_evidence-based_aggregation_method_for_sentiment_analysis
55. Negation Detection Techniques in Sentiment Analysis: A Survey | Iraqi Journal of Science, accessed April 8, 2025, https://ijs.uobaghdad.edu.iq/index.php/eijs/article/view/8680
56. Negation Detection for Sentiment Analysis: A Case Study in Spanish - Simon Fraser University, accessed April 8, 2025, https://www.sfu.ca/~mtaboada/docs/publications/Jimenez-Zafra_etal_NLE_2020.pdf
57. Improving sentiment analysis with multi-task learning of negation | Natural Language Engineering | Cambridge Core, accessed April 8, 2025, https://www.cambridge.org/core/journals/natural-language-engineering/article/improving-sentiment-analysis-with-multitask-learning-of-negation/14EF2B829EC4B8EC29E7C0C5C77B95B0
58. Handling negation and sentiment analysis : r/LanguageTechnology - Reddit,

accessed April 8, 2025, https://www.reddit.com/r/LanguageTechnology/comments/jh2bjt/handling_negation_and_sentiment_analysis/
59. The Power and Precision of Lexicon-Based Sentiment Analysis - LP2M UMA, accessed April 8, 2025, https://lp2m.uma.ac.id/2023/12/27/the-power-and-precision-of-lexicon-based-sentiment-analysis/
60. (PDF) Textual Sentiment Analysis using Lexicon Based Approaches - ResearchGate, accessed April 8, 2025, https://www.researchgate.net/publication/376389784_Textual_Sentiment_Analysis_using_Lexicon_Based_Approaches
61. Sentiment Analysis of News Articles: A Lexicon based Approach | by Farzana huq - Medium, accessed April 8, 2025, https://medium.com/@fhuqtheta/sentiment-analysis-of-news-articles-a-lexicon-based-approach-4672246a12a2
62. An analysis of customer perception using lexicon-based sentiment analysis of Arabic Texts framework - PMC, accessed April 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11154218/
63. Challenges of Sentiment Analysis - Wharton Research Data Services, accessed April 8, 2025, https://wrds-www.wharton.upenn.edu/pages/classroom/challenges-of-sentiment-analysis/
64. An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing - ACL Anthology, accessed April 8, 2025, https://aclanthology.org/W18-4516/
65. Lexicon-Based Sentiment Analysis in Behavioral Research - PMC, accessed April 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11035532/
66. A Comparison of Lexicon-Based and ML-Based Sentiment Analysis: Are There Outlier Words? - arXiv, accessed April 8, 2025, https://arxiv.org/pdf/2311.06221
67. Sentiment Analysis: Techniques, Limitations, and Case Studies in Data Extraction and Classification - International Research Journal, accessed April 8, 2025, https://www.interesjournals.org/articles/sentiment-analysis-techniques-limitations-and-case-studies-in-data-extraction-and-classification-99020.html
68. DHQ: Digital Humanities Quarterly: Sentiment Analysis: Limits and Progress of the Syuzhet Package and Its Lexicons, accessed April 8, 2025, https://www.digitalhumanities.org/dhq/vol/16/2/000612/000612.html
69. What is Sentiment Bias? How will it affect a Lexicon Based Sentiment Analysis?, accessed April 8, 2025, https://datascience.stackexchange.com/questions/63964/what-is-sentiment-bias-how-will-it-affect-a-lexicon-based-sentiment-analysis
70. Sentiment Analysis 101 - Medium, accessed April 8, 2025, https://medium.com/simform-engineering/sentiment-analysis-101-907d4fb6c8fe
71. Sentiment Analysis Challenges: Solutions and Approaches - Determ, accessed April 8, 2025, https://determ.com/blog/sentiment-analysis-challenges-solutions-and-approach

es/
72. An Overview of Lexicon-Based Approach For Sentiment Analysis - IEEC - International Electrical Engineering Conference, accessed April 8, 2025, https://ieec.neduet.edu.pk/2018/Papers_2018/15.pdf
73. Negation detection techniques in sentiment analysis: A survey - Gigvvy Science, accessed April 8, 2025, https://gigvvy.com/journals/ijase/articles/ijase-202306-20-2-003.pdf
74. Scope of Negation Detection in Sentiment Analysis - https ://ris.utwen te.nl, accessed April 8, 2025, https://ris.utwente.nl/ws/files/5513521/DIR_Edited_version_27.pdf
75. A Machine Learning Approach to Negation and Speculation Detection for Sentiment Analysis - ResearchGate, accessed April 8, 2025, https://www.researchgate.net/publication/273600615_A_Machine_Learning_Approach_to_Negation_and_Speculation_Detection_for_Sentiment_Analysis
76. A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis, accessed April 8, 2025, https://www.researchgate.net/publication/333602124_A_Comprehensive_Study_on_Lexicon_Based_Approaches_for_Sentiment_Analysis
77. Sentiment analysis explained 2024 - SuperAnnotate, accessed April 8, 2025, https://www.superannotate.com/blog/sentiment-analysis-explained
78. Comparative analysis of lexicon-based sentiment analysis methods - Sheffield Hallam University Research Archive, accessed April 8, 2025, http://shura.shu.ac.uk/32302/4/Baldwin-ComparativeAnalysisOf%28Pre-print%29.pdf
79. A Stacking Ensemble Based on Lexicon and Machine Learning Methods for the Sentiment Analysis of Tweets - MDPI, accessed April 8, 2025, https://www.mdpi.com/2227-7390/12/21/3405
80. rsher60/Sentiment-Analysis-by-combining-Machine-Learning-and-Lexicon-Based-methods - GitHub, accessed April 8, 2025, https://github.com/rsher60/Sentiment-Analysis-by-combining-Machine-Learning-and-Lexicon-Based-methods
81. Hybrid Framework Integrating Lexicon and Learning Methods for Enhancing Sentiment Analysis Based on Patients' Tweets on Medicine, accessed April 8, 2025, https://journal.esrgroups.org/jes/article/download/3287/2607/5951
82. Combining Lexicon- and Learning-based Approaches for Improved Performance and Convenience in Sentiment Classification - DiVA portal, accessed April 8, 2025, http://www.diva-portal.org/smash/get/diva2:811021/fulltext01.pdf
83. Summing three lexicon based approach methods for sentiment analysis?, accessed April 8, 2025, https://datascience.stackexchange.com/questions/83991/summing-three-lexicon-based-approach-methods-for-sentiment-analysis
84. Taha533/Sentiment-Analysis-of-IMDB-Movie-Reviews - GitHub, accessed April 8, 2025, https://github.com/Taha533/Sentiment-Analysis-of-IMDB-Movie-Reviews
85. SkyThonk/Movie-Reviews-Sentiment-Analysis - GitHub, accessed April 8, 2025, https://github.com/SkyThonk/Movie-Reviews-Sentiment-Analysis

86. Training Dataset for Sentiment Analysis of Movie Reviews - Data Science Stack Exchange, accessed April 8, 2025, https://datascience.stackexchange.com/questions/11220/training-dataset-for-sentiment-analysis-of-movie-reviews
87. IMDb Non-Commercial Datasets, accessed April 8, 2025, https://developer.imdb.com/non-commercial-datasets/
88. Movie Dataset: the 23 Best Data Sets Related to Cinema and TV - Matteo Cassese, accessed April 8, 2025, https://matteoc.com/open-data-entertainment/
89. Q-b1t/IMDB-Dataset-of-50K-Movie-Reviews-Backup - Hugging Face, accessed April 8, 2025, https://huggingface.co/datasets/Q-b1t/IMDB-Dataset-of-50K-Movie-Reviews-Backup