Movie Review Summarization Using Supervised Learning and Graph-Based Ranking Algorithm

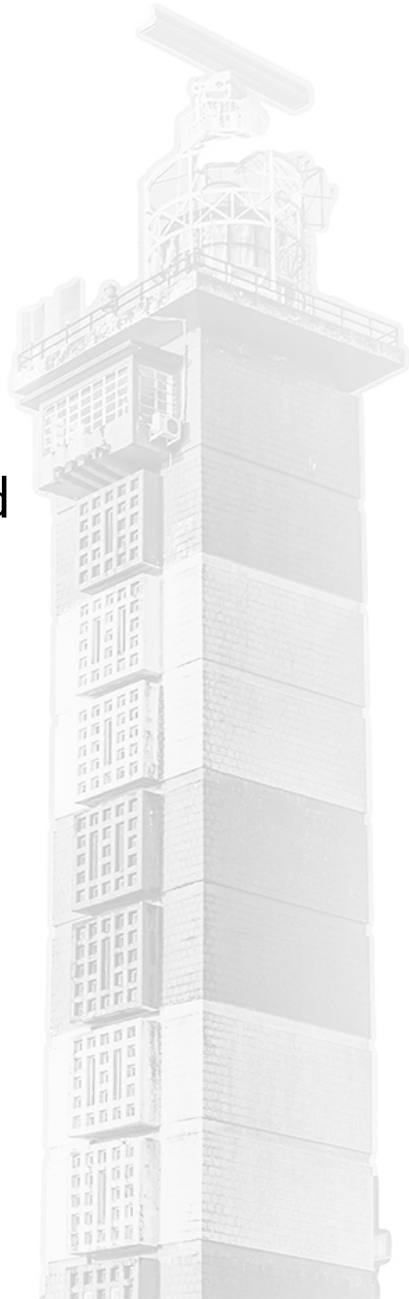
Talapally Sandeep Kumar

Department of Computer Science
National Institute of Technology Karnataka

Mini Project
May 19th, 2022

Contents

- Introduction
- Motivation
- Understanding the Problem Statement
- Objectives and Applications
- Literature Survey
- Dataset Description and Technologies Used
- Plan Of Work
- Data Processing
- Feature Extraction
- Classification Of Reviews
- Review Summarization
- Sentence Embedding
- WGRA Algorithm
- References



Introduction

- Several users post bulk reviews on movie review websites such as IMDB on daily basis.
- These involve user attitude towards a specific movie. Thus, automatically mining and summarizing these bulk reviews is desirable.

Motivation

- The previous approaches proposed for movie summarization are limited to generate feature-based summary rather than generic summary.
- This project proposes a review mining and summarization (RMS) approach that integrates supervised ML approach with graph-based ranking algorithm to automatically generate a generic summary of movie reviews.

Understanding the Problem Statement

- Summarizing thousands of reviews received by a movie can help the viewer (customer) to swiftly scan the summary of it and quickly decide whether to watch a movie or not

Objective and Applications

- The summary of movie reviews can assist the movie service provider such as Netflix to swiftly understand the watching patterns or the interests of their customers.

Literature Survey

- A. F. Alsaqer and S. Sasi, "Movie review summarization and sentiment analysis using rapidminer," in Proceedings of 2017 International Conference on Networks Advances in Computational Technologies (NetACT), pp. 329–335, Trivandrum, India, July 2017.
- A. Trilla and F. Alias, "Sentence-based sentiment analysis for expressive text-to-speech," IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 2, pp. 223–233, 2013.

Dataset Description

- IMDB Dataset: It consists of 50000 movie reviews out of which 25000 are positive reviews and 25000 negative reviews

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive
5	Probably my all-time favorite movie, a story o...	positive
6	I sure would like to see a resurrection of a u...	positive
7	This show was an amazing, fresh & innovative i...	negative
8	Encouraged by the positive comments about this...	negative
9	If you like original gut wrenching laughter yo...	positive

Figure: IMDB DATASET

Technologies Used

- Python 3.8
- Numpy
- Pandas
- NLTK
- SKLEARN
- Tensorflow
- Google Collab

Plan Of Work

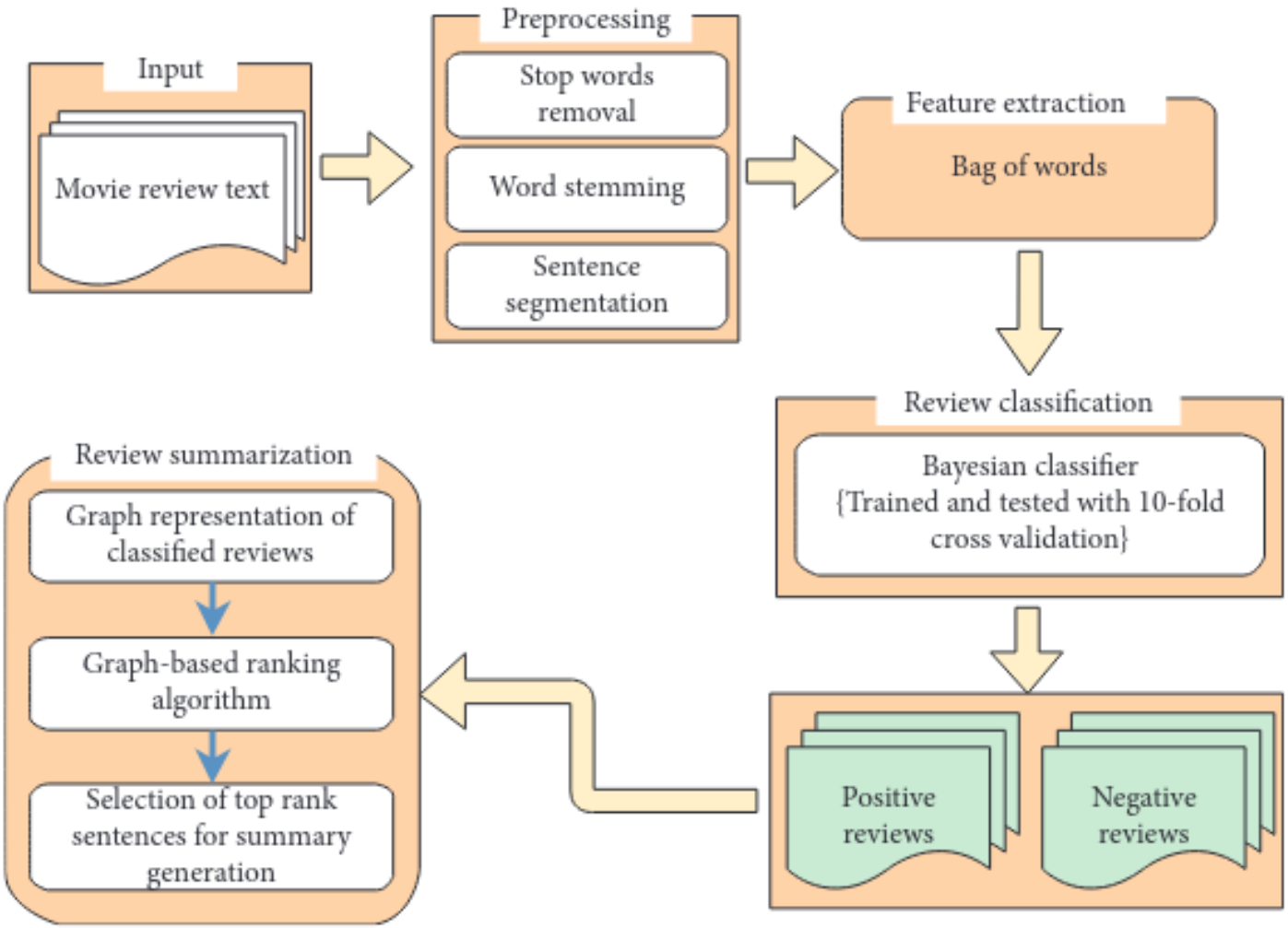


Figure: Solution Approach

Data Processing

- Removal of HTML tags from reviews.
- Removal of Stop Words
- Word Tokenizing
- Performing Lemmatization
- Performing Stemming

Feature Extraction

- This project uses TF-IDF(Term Frequency - Inverse Document Frequency) to extract features for review classification.
- TF-IDF = Term Frequency (TF) x Inverse Document Frequency (IDF)

$$\text{TF-IDF(t,d)} = \text{TF(t,d)} \times \log_{10}\left(\frac{N}{1+df}\right)$$

- Considering both (unigrams and bigrams) for feature extraction.
- Set mindf = 2 and maxdf = 0.5.
- The final output matrix shape has number of reviews as rows and all possible unique words and bigrams as columns

Classification Of Reviews

- In this phase, Multinomial Naive Bayes classification algorithm is used.
- In order to classify the reviews, the feature vectors along with their labels are given as input to the classifier.
- For training and testing of MNB, the 10-fold cross validation technique applied over the given dataset.

Review Summarization

- After classification of given reviews into positive and negative reviews to generate a summary from all the reviews Graph based approach is used to select sentences that are going to be present in the final summary.
- First an embedding for each sentence is created for all the sentences and build a weighted undirected graph G(V,E) where each v_i belongs to V represent a sentence and e_{ij} exists if cosine similarity between v_i and v_j is in range [0,0.5].

Sentence Embedding

- In this project Google's Universal Sentence Encoder(USE) is used for sentence embedding.
- USE has two models for performing sentence embedding one is Transformer model and other is DAN(Deep Averaging Network).In this project DAN model is used.
- The DAN option computes the unigram and bigram embeddings first and then averages them to get a single embedding. This is then passed to a deep neural network to get a final sentence embedding of 512 dimensions.
- It generates a cosine similarity matrix which has shape:(no of sentences , no of sentences).

$$sim(A, B) = \frac{A.B}{||A||.||B||}$$

WEIGHTED GRAPH RANKING ALGORITHM

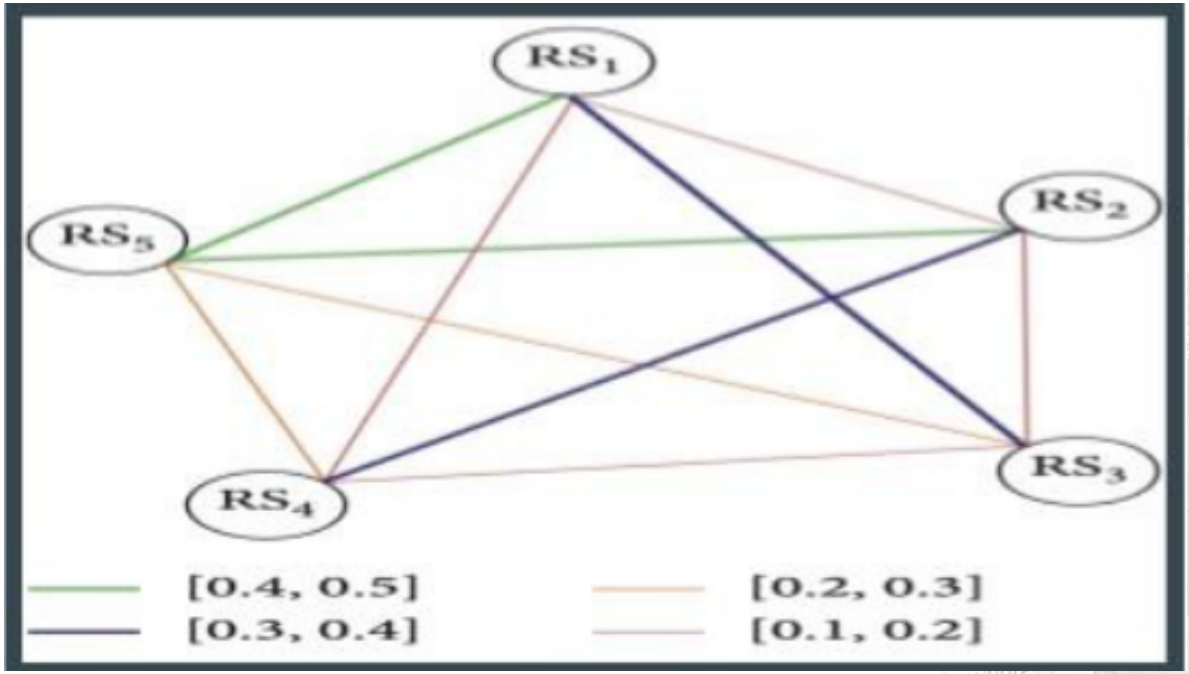


Figure: Undirected Weighted Graph

$$\mathrm{WGRA}(v_i) = (1-d) + d * \sum_{v_j \in \mathrm{In}(v_i)} \frac{\mathrm{WGRA}(v_j) \cdot w_{ji}}{\sum_{v_k \in \mathrm{Out}(v_j)} w_{jk}},$$

Figure: Importance of a node in final summary

Results

```
[ ]  from sklearn import metrics
     predict=model_highest.predict(test_vector)
     accuracy_score = metrics.accuracy_score(predict,
     print(accuracy_score)

     0.8862
```

Figure: Accuracy Of Multi Naive Bayesian Model

```
1 compared to heath's other films like '10 things i hate about you' and even 'brokeback mountain' proves that this man could actually act, doesn't n
2 that's how difficult of a role this was, and that's why ledger's performance is so great.this isn't an action movie.
3 many of the imdb top movies deal with the good old battle between good and evil and the world-weariness of the hero or anti-hero, but this film la
4 it is very well made and gripping with some great performances, but personally i would put casablanca, wizard of oz, amadeus and it's a wonderful
5 in my opinion, the best part of christopher nolan's movies has always been his manipulation of time.
```

Figure: Positive Review Summary of Dark Knight Movie

Results

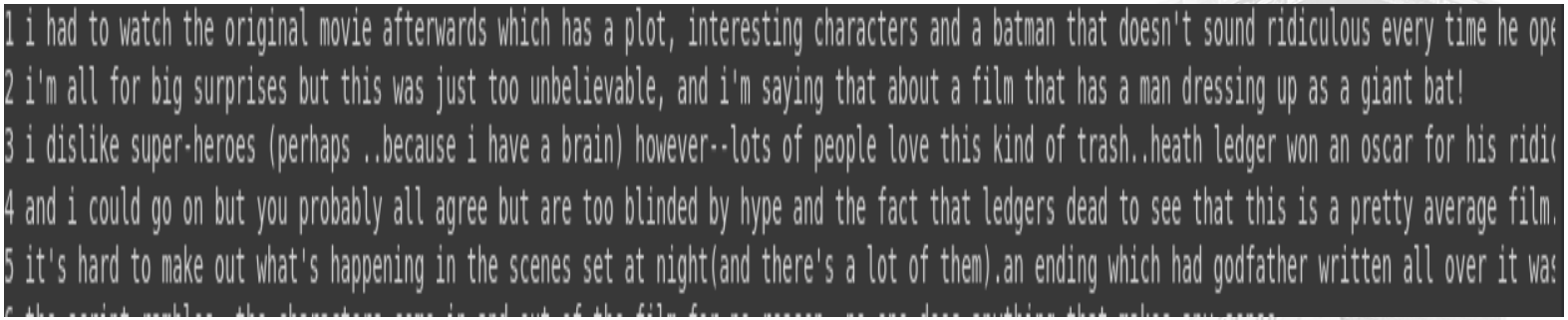


Figure: Negative Review Summary of Dark Knight Movie

References

- Atif Khan and Muhammad Adnan Gul "Movie Review Summarization Using Supervised Learning and Graph-Based Ranking Algorithm "vol 1 ACM,2020
- V. Gupta and G. S. Lehal, "A survey of text summarization extractive techniques," Journal of Emerging Technologies in Web Intelligence, vol. 2, pp. 258–268, 2010.
- .V. B. Raut and D. Londhe, "Survey on opinion mining and summarization of user reviews on web," International Journal of Computer Science and Information Technologies, vol. 5, pp. 1026–1030, 2014.