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

SAI VISHWAK GANGAM(2020202006)
CH N V B DATTATREYA(2020201011
PRUDHVI KOPPURAVURI(2020201010)

International Institute of Information Technology
Hyderabad

30-April-2021

UNDERSTANDING 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.
- 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.

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

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 1: IMDB DATASET

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

SOLUTION APPROACH

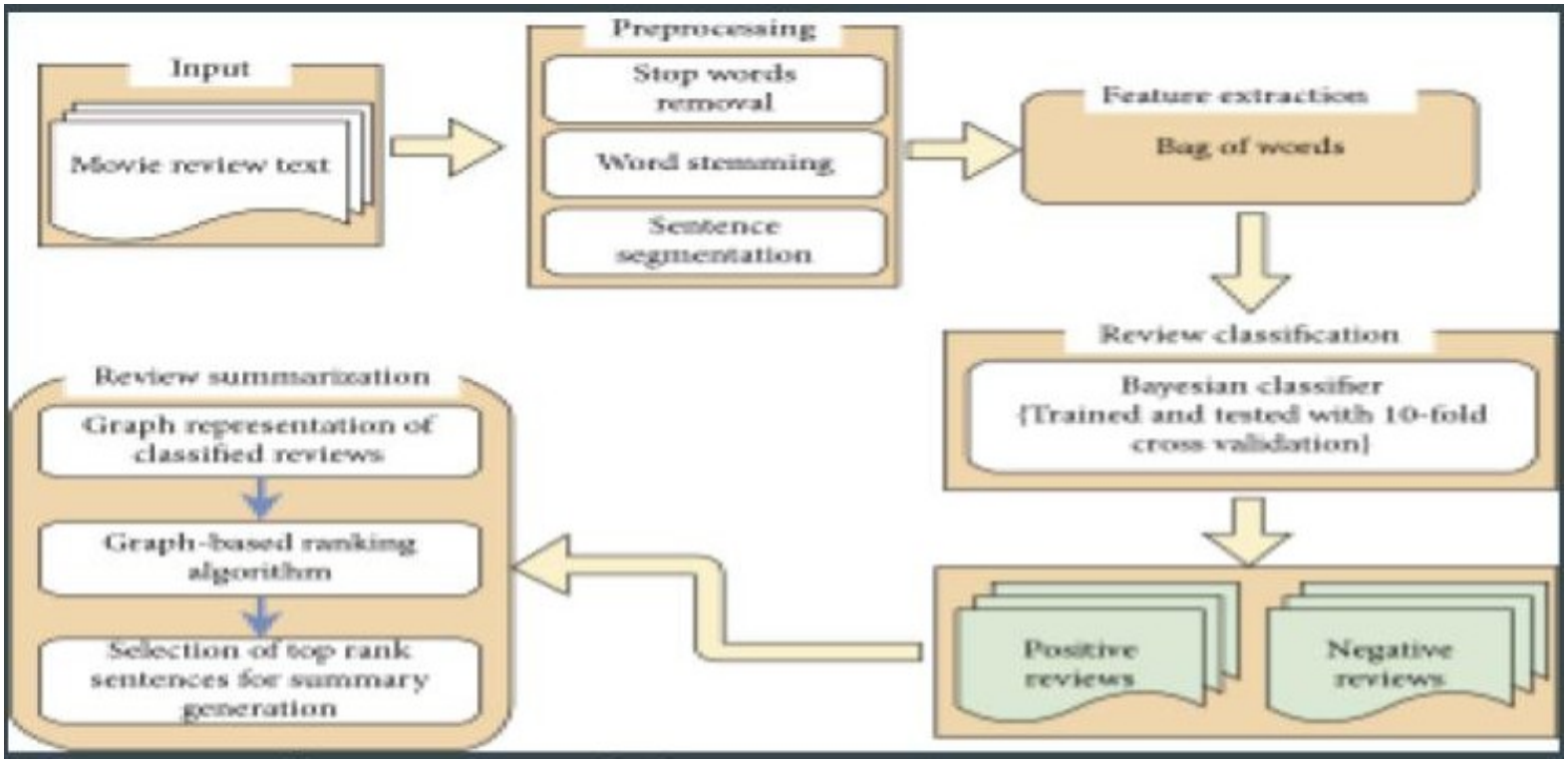

Figure 2: Solution approach

DATA PREPROCESSING

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

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

Feature Extraction

- We used TF-IDF(Term Frequency - Inverse Document Frequency) to extract features for review classification.
- TF-IDF = Term Frequency (TF) * Inverse Document Frequency (IDF)
- We considered both (unigrams and bigrams) for feature extraction.
- We 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

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

CLASSIFICATION OF REVIEWS

- In this phase, we used Multinomial Naive Bayes classification algorithm.
- 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, we applied the 10-fold cross validation technique over the given dataset.

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

REVIEW SUMMARIZATION

- After classification of given reviews into positive and negative reviews to generate a summary from all the reviews we use Graph based approach to select sentences that are going to be present in the final summary.
- First we create a embedding for each sentence for all the sentences and build a weighted undirected graph G(V,E) where each vi belongs to V represent a sentence and eij exists if cosine similarity between vi and vj is in range [0,0.5].

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

SENTENCE EMBEDDING

- We used Google's Universal Sentence Encoder(USE) for sentence embedding.
- USE has two models for performing sentence embedding one is Transformer model and other is DAN(Deep Averaging Network). We used DAN model.
- 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.
- We generated a cosine similarity matrix which has shape:(no of sentences , no of sentences).

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

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

WEIGHTED GRAPH RANKING ALGORITHM

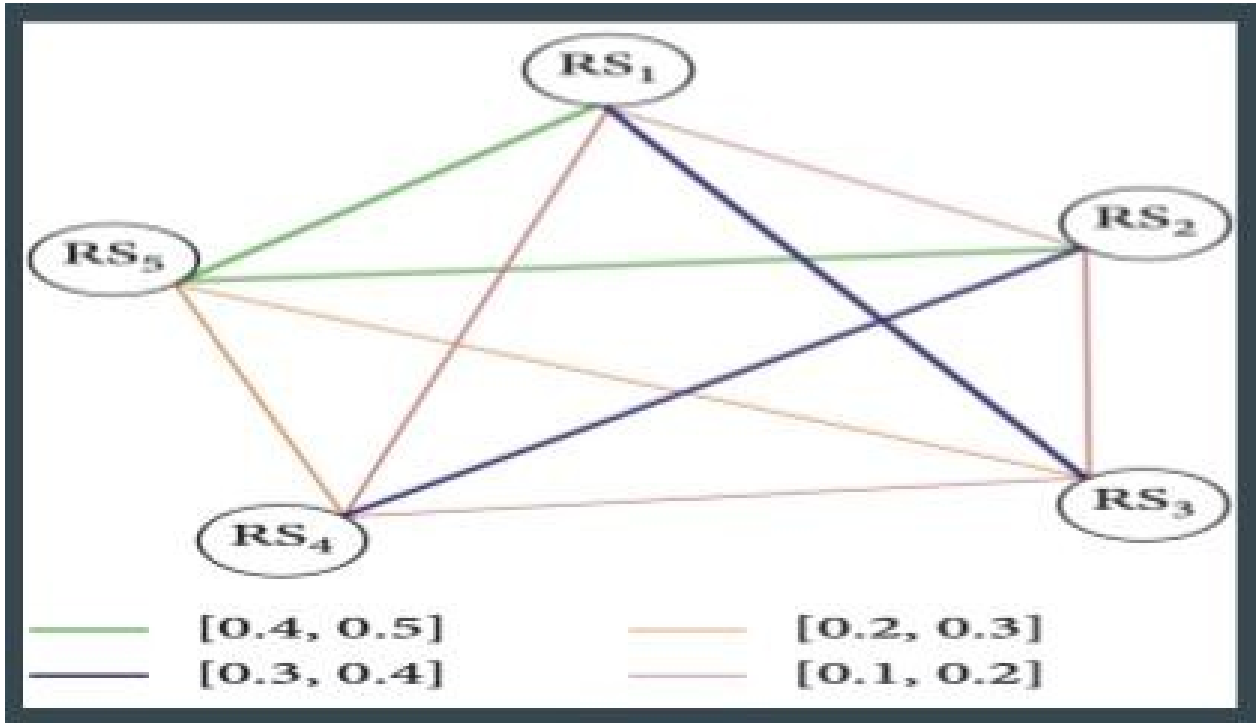

Figure 3: Undirected Weighted Graph

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

Figure 4: Importance of a node in final summary

WORKING DEMO

DEMO

RESULTS

```
1  Even though the film is based on a very romantisised level and not reality, i
   loved it a lot more than the usual biographys or costume drama's.
2  This movie lacks the gravitas and scale to make it a great film, but it's a
   fine cheer-up on a rainy afternoon.
3  I think a lot of the reason i liked the film so much is that the usual silly
   dietrich persona as the "über-vamp" isn't present and her role required her
   to actually act.
4  The atmospherics and the romantic byplay are by far the best part of the
   movie, as viewers are likely to find the resolution a bit of a letdown --
   there's just not that much to it (except a little frisson at the tail end
   that anticipates brian de palma's filmic codas).
5  I was quite prepared to hate the film because of this casting decision, but
   it worked--she was pretty believable and a lot of fun to watch as well!
```

Figure 5: IMDB DATASET POSITIVE SUMMARY

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

RESULTS(contd)

```
1  Well let me tell you something, the movie is not even scary in the least bit.
2  If the most important part of the movie isn't even going to happen, at least
   make it enjoyable to watch and captivating.
3  This was one of the worst films i can remember seeing.
4  I wasn't really disappointed with that matter, but this movie is a matter
   indeed for me, poor plot, useless storyline, naively created and i don't know
   what to say anymore.
5  I'm not saying there should be fighting and crap blowing up but it would
   liven up this more than bland film.
6  Easily one of the worst films ever made.
```

Figure 6: IMDB DATASET NEGATIVE SUMMARY

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

INDIVIDUAL CONTRIBUTION

- PRUDHVI KOPPURAVURI - Data Preprocessing,Web Scraping,Multinomial Naive Bayes
- Sai Vishwak Gangam - TF-IDF,Sentence Embedding using USE
- CH N V B DATTATREYA - WGRA,Review Summarization,Lex Rank

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

References

1."https://www.hindawi.com/journals/cin/2020/7526580/"

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD