Information Retrieval (CS F469)
Project Report

Submitted by:-

Amit Bansal	2016A7PS0140P
Devyash Parihar	2016A7PS0066P
Shubham Lather	2016A7PS0006P

Domain: Sentiment Analysis- Summarization based on multi-view points

---

1. Problem Statement

Opinion Mining from Online Hotel Reviews using Text Summarisation approach.

2. Background of the Problem

a) Description of the selected application domain

The problem in sentiment analysis is classifying the polarity of a given text according to positive or negative sentiments by using the different opinion mining techniques. This problem comes under the domain of Natural Language Processing. To construct a model for Sentiment Analysis, multiple stages are involved in the process. Firstly, the textual data we work on is highly unstructured and hence, pre-processing of data forms a big part of the problem which includes Tokenization, Stop Words removal, POS Tagging etc. After preprocessing, the next stage is the construction of an index. During indexing, algorithms tag sentences based on the polarity and intensity of sentiments. Moreover, conserving the positional semantics of a sentence for indexing perform much better. After all this, the last stage involves the construction of a model which is then used for the classification of new sentences. The Model constructed is able to learn the usage of different terms and their dependence on other terms with which they occur.

b) Motivation of the problem

Nowadays, as there is easy access to the internet and technology, a lot of users are sharing their travel and other experiences about different products on the internet. A major part of this sharing includes the reviews of hotels and places of stay. With the help of these reviews on the internet, future travelers make their decisions of choosing the best places to stay and have a better experience. In today's world, a lot of websites provide a platform for this need of the users, the most popular of them being the TripAdvisor.com. But there is a problem with these websites as they lack an intelligent system which gives the users a summary of reviews for the ease as they provide the complete review which even might not be relevant for the user. And hence, the user has to scroll through many reviews to understand the sentiments of the majority. For an intelligent system, we need to take into account many factors like the credibility of the reviewer, time for which a review is posted, the seriousness of review, different opinions of different authors etc. to generate an extractive summary of the reviews.

c) Technical issues included

Bag of words, cosine similarity calculation and Posting Lists.

3. Literature Survey

Sentiment analysis is an application of natural language processing also known as emotion extraction or opinion mining. The basic idea is to find the polarity of the text and classify it into positive, negative or neutral. Having seen a lot of developments in many different applications, it has also found its use in finding the sentiments of social media users and the polarity of product and movie reviews which help humans in making their decisions.

Effective identification of hotel reviews has become an important issue as there is information overload due to the expanding amount of online hotel reviews. In [1], the authors have tried to overcome the major challenge of sorting the reviews by publication date or by the number of votes. Recently published reviews may have a lower number of votes or vice versa. So, to identify the helpful reviews, they investigated the helpfulness of these reviews from the aspects of review quality, review sentiment and also the reviewer characteristics. They used various classification techniques to develop review helpfulness prediction models. To identify a review as helpful, they assumed that the customer has actually read the review and identifies it as helpful and it can provide valuable information which further affects a customer's decision. In this method, the authors preprocessed the reviews including calculation of review lengths, words and sentence segmentation and part-of-speech tagging. After this, they calculated the helpfulness by using the helpful votes for review and the number of days elapsed till review was crawled. To construct the classification model, classification techniques used were Decision Tree, Random Forest, Logistic Regression and Support Vector Machine. Then to calculate the model performance, they constructed the metrics of accuracy, sensitivity, specificity, precision, recall, and F-measure.

The authors of [2] have introduced the QMOS method, which employs a combination of sentiment analysis and summarization approaches. It a lexicon-based method for query-based multi-document summarisation for opinions expressed in user reviews. The proposed approach takes into consideration the contextual polarity of review sentences, includes sarcasm detection and handles subjective sentences polarity. They've then used a graph-based model for ranking and extractive summarization of sentences with respect to the query. The edges between sentences and query represent a similarity score between them. The method employs the greedy algorithm and query expansion approach to reduce redundancy and bridge the lexical gaps for similar contexts that are expressed using the different wording, respectively. In this, the authors conclude that the QMOS method can significantly improve the performance and make QMOS comparable to other existing methods.

The authors of [3] talk about an approach for prediction of customer opinion using supervised machine learning approach and Decision Tree method for classification of online hotel reviews as positive or negative. In this methodology, unstructured data is

turned into a structured format and text data is preprocessed which includes a transformation to lower case, splitting into a sequence of tokens, removal of stopwords. Finally stemming is performed by Porter stemming algorithm to reduce to a minimum length of the word after which TF-IDF score is calculated. The result from preprocessing is in the form of a term-document matrix, where each token is now an attribute in a column and each review is an example in a row and cells contain TF-IDF scores which are used by the classifier. The decision tree algorithm incorporates attribute selection by using Information gain as a criterion for evaluation of attribute importance; during preprocessing, applying TF-IDF diminishes the weight of terms that occur very frequently in the data set and increases the weight of terms that occur rarely. And then decision tree algorithm is used to generate a classification model for predicting the values of a target attribute (class or label) based on the values of several input attributes in the training data, used for classification of reviews. The prediction accuracy of the model is evaluated using k fold Cross-validation. Cross-validation divides the training data set into 10 equally sized, non-overlapping subsets and the model is trained on the first nine sets and tested on the tenth remaining set.

In [4], authors address three hidden assumptions prevalent in online review studies which are: all reviews are visible equally to online users, review rating(RR) and hotel star class(HSC) affect review helpfulness individually with no interaction and characteristics of reviews and reviewer status stay constant. After collection of data, the reviews without HSC & RR and the reviews that had received no helpfulness vote are removed. In this study, review helpfulness is defined as the number of users who have voted the review as helpful. The independent variables are divided into four categories: the review content (CO), sentiment (SE), author (AU), and visibility (VI) features. In this analysis, linear regression to a piece of text is applied and the degree of education level required for a reader to comfortably grasp what was written in the review is calculated using different methods. Stanford POS tagger is used to tag the words. After that, subjectivity and polarity classifiers were performed to identify sentence subjectivity and word polarity, respectively. Three Data mining techniques including a linear regression, model tree and support vector regression are used. Prior to model construction, a correlation-based feature selection (CFS) method was used to evaluate the correlations between the feature subsets and the dependent variable. In this study, the greedy stepwise algorithm is used for CFS procedure and ten-fold cross-validation technique is adopted to build the training and testing data sets.

The authors of [5] have enumerated and analyzed various the various approaches which are followed for sentiment analysis. Two machine learning methods Naive Bayes Classifier and Support Vector Machines which generate more improved and precise result as compared to earlier techniques like human generated baselines were used. As comparing in between both the techniques of machine learning SVMs results are better

than Naive Bayes Classifier, due to their inability to work on linearly inseparable data, although differences weren't very large. After SVMs, to achieve extremely high accuracies, Neural Networks and their variations are being used.

In [6], the authors have analyzed different summarisation systems. They examine the role of three principal components of an extractive summarisation technique: sentence ranking algorithm, sentence similarity metric, and text representation scheme. Three broad categories of extractive summarization systems are centrality based, corpus-based and graph based. Centrality based system find centra sentences whose similar sentences are included in the summary. Graph-based system constitutes sentences as node and similarity as an edge. Summaries are obtained by looking at the number of nodes connected to a given node. both of these techniques rely on the corpus of documents and iteratively select sentences to be included in the summary. The authors state that most state-of-art systems can be traced back to simpler systems. Two of these systems are:

Greedy-KL: It follows a greedy algorithm to minimize the KL Divergence between the original document and resultant summary.

Centroid: After calculating the centroid, sentences that are close to it in terms of cosine similarity are iteratively chosen.

In this, authors have primarily highlighted the fact that for a new text summarization technique, the importance of finding a better sentence similarity measure is more than sentence ranking algorithm.

In [7], the authors have proposed an approach of opinion mining from online hotel reviews using a novel multi-text summarization technique for identifying the top-k most informative sentences of hotel reviews. Initially, they build a vocabulary which consists of only nouns, adjectives, and adverbs as only these words carry semantically distinguishing features for sentiment analysis. After this as compared to previous studies on review summarization, this approach has considered critical factors like author credibility, review time, review usefulness and conflicting opinions of different authors. And using these factors, metrics for sentence importance calculation are constructed which is then used to represent a sentence. Metrics for sentence similarity is also created considering content and sentiment based similarity. And then to identify the top-k sentences, they ranked the sentences using the k-medoid algorithm. For ranking, they first clustered the data into p clusters. A representative sentence is chosen from each cluster and then their cosine similarity is calculated after which top k of these p sentences are chosen.

The authors of [8] have talked about the Fuzzy logic Extraction approach for text summarization and semantic approach of text summarization using the recent semantic analysis. After preprocessing, each sentence is represented by the attribute of a vector of 8 features who has a value 0 or 1. After extraction of 8 features, the result is passed to fuzzifier then to inference engine and finally to defuzzifier. Rules for inference engine are

supplied from Fuzzy rule base after which sentence will be sorted in decreasing order of score calculated from the fuzzy logic method. And then, sentences with the highest score are extracted as a document summary. Authors also used Latest semantic model which compares the semantic similarity between pieces of textual information which includes three steps: Input Matrix Creation, Singular Value Decomposition and Sentence Selection. After using these two approaches, the final set of an improved summary is obtained by the union of summaries obtained from each approach. In the end, authors conclude that incorporating the latent semantic analysis into the sentence feature extracted the fuzzy logic system to extract the semantic relations between concepts in the original text improves the quality of the summary.

As most of the reviews are written in an informal language and usually do not follow standard language rules, introducing a lot of noise in the process of text summarization. To overcome this, the authors of [9] present their investigations to generate extractive and abstractive summaries of opinions. They evaluate the previously used methods according to three measures: informativeness, linguistics quality and utility of the summary. In extractive summarization method, summaries are generated in such a way that aspect coverage and the distribution of polarity become preserved as much as possible using sentence clustering and sentence ranking. In the abstractive summarization method, more natural summaries are generated. Some approaches for this are the sentence fusion, sentence compression etc. This method has two phases: clustering of textual segments and text generation based on templates. In this paper, authors make the following contributions: a new content selection strategy to produce extractive summaries of reviews, a novel Natural Language Generation (NLG) template-based system to generate abstractive summaries of opinions, and comparison of extractive and abstractive opinion summarization methods. In general, the extractive method obtained the best results as compared to the abstractive method. In relation to informativeness and utility of summaries, extractive methods are better, however, in relation to the non-redundancy criterion, abstractive method outperforms the extractive method.

A term based method cannot deal with the problems of polysemy and synonymy while ontology-based method requires lots of manpower but it takes into account the semantic information of document content. To overcome these problems, authors of [10] presents a pattern based model for generic multi-document summarization, which exploits closed patterns to extract the most salient sentences from a document collection and reduce redundancy in the summary. This method combines the advantages of the term-based and ontology-based models while avoiding their weaknesses and significantly outperforms the state-of-the-art methods.

4. Research Gap

- An approach can be proposed which also takes into account the sarcasm detection and handles subjective sentences polarity.
- An approach which takes into account the visibility of all reviews equally for all users could also be used so that there is an equal chance for each review to be get voted by a user.
- As different users can use some words which require a wide knowledge of a language (English in this case), an approach, which can calculate the degree of education level required for a reader to comfortably grasp what was written in the review, can be introduced.
- As most of the reviews written by the users are written in an informal language and they usually do not follow the standard language rules, some extractive or abstractive summaries methods can be used. Also, linguistics quality can be taken into account.
- To reduce the redundancy and to deal with the problems of polysemy and synonymy, a new approach could have been introduced.

5. System Description

First, the reviews were broken down into sentences by splitting the reviews along period(.). Then the stopwords were removed and POS tagging was done. All this was put in a data frame which retained the original review ID. As the paper suggested, only nouns, adjectives and adverbs were retained. Then the following measures were developed to calculate the importance of similarity of sentences.

1. **Author credibility:** It is calculated as the mean error between the rating given by the author and the average rating of the hotel.

2. **Author recommendation score:** It is based on the number of likes the comments of an author receive.

3. **Representativeness of review author:** It is the average of Author credibility and Author recommendation score.

4. **Review recency:** It takes into account when the review was written and when the query was made.

5. **Review sentence score:** It assigns a score to each sentence based on its location (more weight if it's the first sentence in the review) if it contains indicator phrase and the length of the sentence.

6. **Sentence Importance:** It is the average of the (3) and (4) multiplied by (5).

7. **Sentiment Similarity:** Semantic Orientation-Pointwise Mutual Information (SOPMI) method was used to calculate the sentiment score of the sentences. SOPMI gives a score to every adjective by the difference of the number of associations of that particular adjective with positive adjectives and negative adjectives. Positive SOPMI score implies positive sentiment orientation of the word and vice versa.
   Then SOPMI score for all adjectives in a sentence was averaged.
   This average was used to find the polarity of the sentence.
   Then this polarity of different pairs of sentences was compared to assign a sentiment similarity to a pair of sentences.

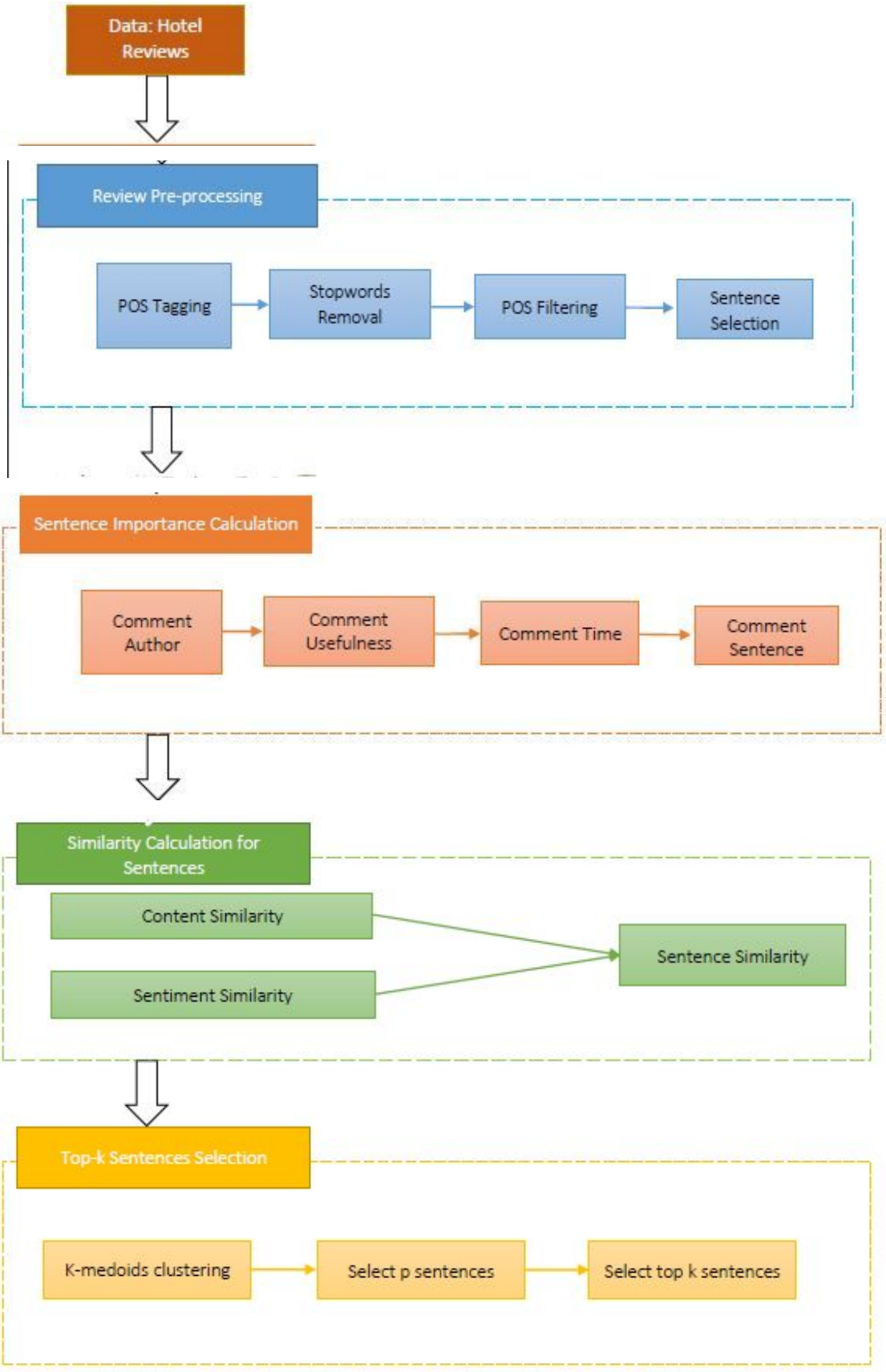
8. **Sentence Similarity:** Finally, the cosine similarity of the sentences was multiplied by the sentiment similarity to find the sentence similarity.

   With these measures ready, clustering was done in an effort to find the representative sentences. The k-medoids algorithm was used to make p number of clusters, where p is the number. The sentence similarity measure calculated earlier was used to cluster most similar sentences together. Out of these p clusters the most representative sentence was found and out of these the top k sentences were given as the result.

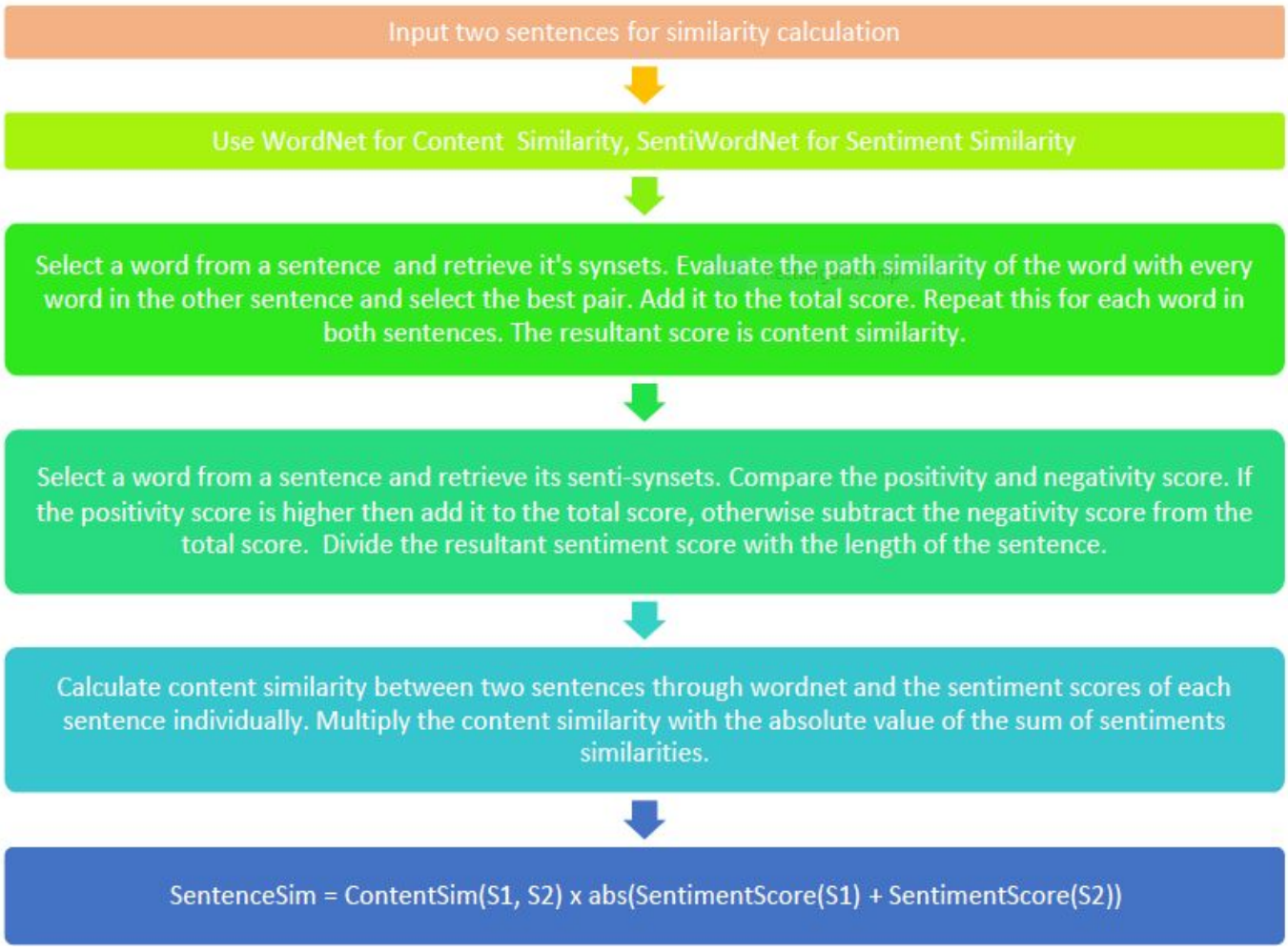
Some of the measures suggested by the authors were not possible to implement:

1. We do not have the required data to calculate the author recommendation score as it would require the number of likes the review has acquired from other users.

2. The authors have calculated sentence similarity as the product of content similarity with sentiment similarity. To calculate content similarity, Normalized Google Distance has been used. This is a paid API by Google and it is not accessible to us. Hence cosine similarity has been used in place of Content Similarity.


Data: Hotel Reviews
Review Pre-processing
POS Tagging
Stopwords Removal
POS Filtering
Sentence Selection
Sentence Importance Calculation
Comment Author
Comment Usefulness
Comment Time
Comment Sentence
Similarity Calculation for Sentences
Content Similarity
Sentiment Similarity
Sentence Similarity
Top-k Sentences Selection
K-medoids clustering
Select p sentences
Select top k sentences

Improvement in the paper:



We have tried to improve upon this calculation of similarity as it is one of the most important aspects of the method. We have used Cosine Similarity to calculate how similar two words are.

We have borrowed the idea of sentence similarity being a product of content similarity and sentiment similarity. We have used the SentiWordNet corpus for capturing sentiment similarity.

The SentiWordNet Corpus consists of words stored as senti-synsets. Each senti-synset has 3 attributes:

- Positivity Score
- Negativity Score
- Objectivity Score

The sum total of the above-mentioned score is 1. In our approach, we have used only positivity and negativity scores. This is because we already filter out the words which have low objectivity during the POS-filtering stage.
The two scores of each pos-tagged word are retrieved from the senti-wordnet corpus. The score having higher absolute value is selected. If the word is positive (i.e. it has a higher positive score), the positive score is added to the total sentence score. If the word is negative, the negative score is subtracted from the total sentence score. If the total sentence score is positive, then it is a positive sentence and vice versa. The scores of two sentences being compared are added and the absolute value is taken. There can be three cases:

- Both sentences are positive: The scores will add up to give a higher positive value.

- Both sentences are negative: The scores will add up to give a more negative value and therefore a high absolute value.

- One sentence is positive, one is negative: The scores will cancel each other and produce a relatively low absolute value indicating low sentence sentiment similarity.

6. Evaluation Strategy

We have used the TripAdvisor dataset which is arranged in JSON format. The data has the rating on multiple criteria with the most important being Overall Rating. The data also have the time when the review was written, author name, hotel name etc. The data has 12,773 JSON files, each having the review of one hotel. For testing purposes, we have used a small portion of the data.

For clustering three types of similarity have been used:

1. **Cosine similarity:** Using cosine similarity to decide how similar two sentences are.

2. **Cosine Similarity with SOPMI:** We multiply the cosine similarity with the SOPMI score to give more weightage to the sentiment similarity to help cluster sentences with similar sentiments together.

3. **Cosine Similarity with SentiWordNet[WordSent]:** Using the synsets from SentiWordNet, we have calculated sentence similarity. We find one similarity score with Cosine Similarity which pertains to content similarity. The sentiment scores for each sentence are calculated using SentiWordNet and then the absolute value of their sum is multiplied with the Content Similarity.

7. Experimental Results and Evaluation

Hotel ID	Ranking	Cosine Similiarity	SOPMI x Cosine Similarity	WordSent
73743	1	Aside from the bad side of town where this hotel is located, the obnoxiously rude staff, the offensive odor of manure in the hallways, the difficulty of opening an obviously non-serviceable door to our room, dirty carpets, sticky fingerprints on phones, remotes for two TV's nowhere in sight (not to mention the staff's unwillingness to remedy the problem), dirty bathtub, broken screen door to the balcony, missing vents on outdated AC/heater unit, missing smoke detectors, hair on pillowcases, stiff mattresses, constant scent of ashtrays in room	We stayed in this inn with my friends overnight, bed was comfortable, room was clean and spacious	Dad spent two nights in one room at the Best Western with three kids 18 to 24 years old
	2	Don't listen to anything these other people are saying	The reviews on this hotel are right on the money	The Best Western Plus Executive Inn is a nice clean hotel located within walking distance of many Seattle attractions (Seattle Space Needle, Ride the Ducks, Seattle waterfront, Pike Place Market, the mono-rail)

	3	Front desk staff was on the verge of unfriendly at check in (night staff was very nice)	A lot of money has been sunk into this hotel since Victory took it over from Comfort Inn	The restaurant had a perfect breakfast buffet, and the parking was free
72572	1	I had a large, clean room located in the middle of the hotel	We were happy to find this hotel in historic Pioneer Square	I stayed at this hotel pre cruise,the location is great, next to the Space Needle and the monorail which takes you into the centre of the city, then a quick walk and you are at Pikes market which is well worth a visit
	2	We were happy to find this hotel in historic Pioneer Square	Great hotel, great service,very nice breakfast, the rooms are beautiful, love the charm and character of the building and neighborhood	I was looking for a place for my wife and I and our two teenage daughters to stay before embarking on our cruise from Seattle to Alaska
	3	Very friendly and helpful staff, good complimentary breakfast	The staff were very friendly and helpful	The staff was very patient with my concerns about my credit card
	1	I stayed at this hotel pre cruise,the location is great, next to the Space Needle and the monorail which takes you into the centre of the city, then a quick walk	I stayed at this hotel pre cruise,the location is great, next to the Space Needle and the	The staff was very patient with my concerns about my credit card

		and you are at Pikes market which is well worth a visit	monorail which takes you into the centre of the city, then a quick walk and you are at Pikes market which is well worth a visit	
72586	2	Nice motel, clean, rooms are warm and smell good	Dad spent two nights in one room at the Best Western with three kids 18 to 24 years old	I stayed at this hotel pre cruise,the location is great, next to the Space Needle and the monorail which takes you into the centre of the city, then a quick walk and you are at Pikes market which is well worth a visit
	3	We recieved a voucher for a free night stay in a Best Western from a BW summer promotion	We recieved a voucher for a free night stay in a Best Western from a BW summer promotion	Dad spent two nights in one room at the Best Western with three kids 18 to 24 years old

8. Conclusion and Future Work

The above table encapsulates the results for three different approaches. The Cosine Similarity and SOPMI approach are suggested by the paper. WordSent approach is the improvisation on their approach. While assessing the clusters, it was seen that all 3 approaches well-segregated sentences based on different topics.
It can also be seen that our approach clubs together content similarity of various sentences and the most encompassing sentence comes out on the top. In the case of hotel ID 72572, the top sentence for SOPMI is 'We were happy to find this hotel in historic Pioneer Square' which is good but not very informative. On the other hand, we have a similar top review but much more informative in case of WordSent: 'I stayed at this hotel pre cruise,the location is great, next to the Space Needle and the monorail which takes you into the centre of the city, then a quick walk and you are at Pikes market which is well worth a visit'.

There is a lot of plausible future work for the said application.

1. Models such as word2vec can be used to capture the semantic occurrence and similarity of sentences. Cosine similarity can then be used to model sentence similarity.

2. A better clustering algorithm than k-medoids can be used. Algorithms such as Chameleon help capture the relative similarity and dissimilarity between sentences in forming clusters.

3. Verbs and degree adverbs (such as like, love, hate, etc.) can be included in future studies to see their effect in capturing sentence sentiments.

9. References

1. Pei-Ju Lee, Ya-Han Hu, and Kuan-Ting Lu. “Assessing the helpfulness of online hotel reviews: A classification based approach”. Telematics and Informatics 35 (2018) : 436-445.
2. Asad Abdi, Siti Mariyam Shamsuddin, and Ramiz M. Aliguliyev. "QMOS: Query-based multi-documents opinion-oriented summarization." Information Processing & Management 54.2 (2018): 318-338.
3. Stanimira Yordanova and Dorina Kabakchieva. "Sentiment Classification of Hotel Reviews in Social Media with Decision Tree Learning". International Journal of Computer Applications (0975-8887): Volume 158- No 5, January 2017.
4. Ya-Han Hu, Kuanchin Chen. "Predicting Hotel Review Helpfulness: The impact of Review Visibility, and interaction between Hotel Stars and review ratings". International Journal of Information Management 36 (2016): 929-944.
5. Gupta, Pankaj, Ritu Tiwari, and Nirmal Robert. "Sentiment analysis and text summarization of online reviews: A survey." Communication and Signal Processing (ICCSP), 2016 International Conference on. IEEE, 2016.
6. Mehta, Parth, and Prasenjit Majumder. "Effective aggregation of various summarization techniques." Information Processing & Management 54.2 (2018): 145-158.
7. Hu Ya-Han, Yen-Liang Chen, and Hui-Ling Chou. "Opinion mining from online hotel reviews–A text summarization approach." Information Processing & Management 53.2 (2017): 436-449.
8. S.A. Babar, Pallavi D.Patil. "Improving Performance of Text Summarization". Procedia Computer Science 46 (2015) 354-363.
9. Roque Enrique Lopez Condori, Thiago Alexandre Salgueira Pardo. "Opinion summarization methods: Comparing and extending extractive and abstractive approaches". Expert Systems with Applications 78 (2017): 124-134.
10. Ji-Peng Qiang, Ping Chen, Wei Ding, Fei Xie, Xindong Wul. "Multi-document summarization using closed patterns." Knowledge-Based Systems 99 (2016): 28-38.