Read what you need: Controllable Aspect-based Opinion Summarization of Tourist Reviews

Rajdeep Mukherjee  
IIT Kharagpur, India  
rajdeep1989@iitkgp.ac.in

Hari Chandana Peruri  
IIT Kharagpur, India  
chandupvsl@iitkgp.ac.in

Uppada Vishnu  
IIT Kharagpur, India  
vishnu14july@iitkgp.ac.in

Pawan Goyal  
IIT Kharagpur, India  
pawang@cse.iitkgp.ac.in

Sourangshu Bhattacharya  
IIT Kharagpur, India  
sourangshu@cse.iitkgp.ac.in

Niloy Ganguly  
IIT Kharagpur, India  
niloy@cse.iitkgp.ac.in

ABSTRACT
Manually extracting relevant aspects and opinions from large volumes of user-generated text is a time-consuming process. Summaries, on the other hand, help readers with limited time budgets to quickly consume the key ideas from the data. State-of-the-art approaches for multi-document summarization, however, do not consider user preferences while generating summaries. In this work, we argue the need and propose a solution for generating personalized aspect-based opinion summaries from large collections of online tourist reviews. We let our readers decide and control several attributes of the summary such as the length and specific aspects of interest among others. Specifically, we take an unsupervised approach to extract coherent aspects from tourist reviews posted on TripAdvisor. We then propose an Integer Linear Programming (ILP) based extractive technique to select an informative subset of opinions around the identified aspects while respecting the user-specified values for various control parameters. Finally, we evaluate and compare our summaries using crowdsourcing and ROUGE-based metrics and obtain competitive results.

CCS CONCEPTS
• Information systems → Personalization; Summarization; Information extraction; Sentiment analysis.

KEYWORDS
Controllable summarization, Aspect-based opinion mining, Unsupervised extractive opinion summarization, Personalization, Tourism

1 INTRODUCTION
Tourism and Hospitality is an ever-growing industry and a crucial driver for the economic growth and development of any nation. Every year, millions of tourists travel across the world and share their pleasant/unpleasant experiences online through various platforms in the form of blogs and reviews. Future travelers and local service providers are benefited alike from this wide range of opinionated information for shaping their decisions [13]. However, manually going through all reviews and extracting relevant knowledge from them is an overwhelming task. Also, readers may have different levels of information needs which are not considered by the current state-of-the-art approaches for multi-document summarization [5, 14]. We propose an unsupervised and controllable summarization framework which extracts aspect-based opinion summaries from huge corpora of tourist reviews, the shape and content of which can be customized according to reader’s preferences.

Opinion Summarization, as introduced in [11], deals with (1) finding relevant aspects (such as amenities, culture, etc.) about the entity being reviewed (here, place) and identifying sentences which discuss them; (2) discovering sentiments expressed towards the identified aspects; and (3) generating a concise and digestible summary of opinions. While general-purpose summarization approaches [5, 14] try to optimize the salience of the overall generated content, opinion summarization [6] is more challenging as it is aspect-centric and focuses on maximizing the diversity of opinions being covered in the final summary. Supervised methods for opinion summarization such as [18] depend on large annotated datasets of document-summary pairs to train their models, which make them difficult to adapt across domains [4]. Motivated by recent works such as [3, 21], we propose an unsupervised extractive summarization framework built on top of an unsupervised aspect extraction module which makes our method generalizable to any domain.

Travel reviews posted online capture a wide range of human emotions owing to reviewers’ social and cultural backgrounds. State-of-the-art methods for opinion summarization [4, 6] do not allow for personalization while generating summaries from such diverse range of opinionated text. We, however, argue that readers should be able to customize the shape and content of such summaries to suit their varying interests and time budgets. To our knowledge, only [2, 8] have motivated the need for controllable summarization. However, both of them are supervised techniques and require gold-standard summaries for training their models. Our proposed framework is unsupervised and extractive in nature which additionally lets our readers control several attributes of the summary.
such as its length and specific aspects of interest it must focus on. Motivated by [7], our default summaries contain an equal proportion of opinions from male and female reviewers. We further make this ratio customizable to suit user preferences.

For our experiments, we create a dataset consisting of user reviews posted on TripAdvisor (refer Section 6). First, we identify relevant aspects from the data following an unsupervised approach as proposed in [10] (refer Section 4). Then, we assign a score to each review sentence based on its salience. Finally, our ILP-based summarization framework extracts an optimal subset of sentences that best summarizes opinions about various aspects of the place being reviewed (refer Section 5). We achieve competitive performance on a variety of evaluation measures when compared against our baseline methods for unsupervised summarization (refer Section 6). We make our code, datasets, survey instruments and a link to our web application publicly available at https://github.com/rajdeep345/ControllableSumm/.

2 RELATED WORK

We do not propose a new method for aspect extraction. However, as motivated in Section 4, we need to limit the number of aspect options, a control parameter in our proposed framework. Dependence of supervised methods such as [19] on considerable amount of labeled data and their inability to cluster the extracted aspect terms into coherent topics make them unsuitable for our purpose. We, therefore, follow an unsupervised approach as proposed in [10] to identify relevant aspects from our tourist-review corpus.

Supervised methods for opinion summarization [6] depend on significant volumes of training data which is difficult to procure across domains [4, 5]. Our ILP-based summarization framework, on the other hand, draws its motivation from unsupervised approaches such as [3, 17, 21]. None of these methods, however, facilitate readers to decide and control what they read. Though [2, 8] take user preferences into account, they are supervised techniques and require gold-standard summaries for training. Our proposed framework is unsupervised and extractive in nature. It further allows readers to control the shape and content of the generated summaries according to their preferences.

3 PROBLEM FORMULATION

Let $R_p = \{R_{p1}^p, R_{p2}^p, ..., R_{pn}^p\}$ represent the set of $n$ reviews available for a place $p$, where each review $R_{pj}^p = \{R_{p1j}, R_{p2j}, ..., R_{pmlj}\}$ represents a set of $m$ sentences. $S_p = \bigcup_{i=1}^n \bigcup_{j=1}^m R_{ij}$ therefore represents the set of all review sentences available for the place. Let, the set of all aspects, as identified by our unsupervised aspect extraction module, be represented by $A = \{A_1, A_2, A_3, ..., A_K\}$. In our proposed solution for controllable summarization, readers have the flexibility of selecting the aspects they are interested in. Let that set be represented by $A' \subseteq A$. Further, let $L$ (words) represent the desired length of summary. The aim of our ILP-based summarization module is to select a subset $S' \subseteq S_p$ which best summarizes the reviews in $R_p$ within $L$ words while capturing the important opinions expressed towards the aspects in $A'$. As all users may not always be interested in reading fair summaries, hence the obtained summary should additionally take into account the desired ratio of comments from female and male reviewers as set by the reader.

4 ASPECT IDENTIFICATION

For the task of identifying relevant aspects, we follow an unsupervised Attention-based Aspect Extraction (ABAE) technique as proposed in [10]. ABAE is essentially an autoencoder-based topic model where the goal is to learn a set of $K$ aspect embeddings without any supervision, where $K$ represents the number of topics/aspects to be identified. Given a set of review sentences as input, the model is trained to minimize the sentence reconstruction loss while learning to attend on the aspect words (please refer [10] for details). Finally, each extracted topic is manually interpreted by looking at its representative words and assigned a genuine aspect label.

In order to quantify the interpretability of each identified topic cluster, we calculate its coherence score as defined in [10]. The value of $K$, for which the average coherence score of all the clusters is maximum, is selected as the ideal no. of discoverable topics from the review corpus. However, based on the results of a survey with 30 participants (frequent travelers and interested in reading opinion summaries of tourist reviews), we find the ideal value of $K$ obtained in our experiments, to be much higher than the ideal no. of aspects readers would like to choose from while generating the summaries. This establishes the need for clustering the identified fine-grained topics into coarse-grained aspect classes (Refer Section 6 for results).

5 SUMMARIZATION FRAMEWORK

5.1 Opinion Scoring

As a first step of opinion summarization, we assign a salience score to each review sentence based on the following three criteria:

**Readability Score:** In order to capture easily comprehensible opinions, we measure the readability of a sentence by means of its Flesch Reading Ease score (0=very difficult and 100=very easy), which has been used as a widely-accepted metric since many years [20] for evaluating the simplicity of language.

**Sentiment Strength:** We use the CoreNLP SentimentAnnotator [12] to obtain the sentiment polarity score of a sentence. It is an integer value in [0, 4] range, with 0 and 4 representing the most negative and positive sentiments, respectively. We obtain the sentiment strength as an absolute difference of this score with 2.

**Aspect-Relevance:** Each sentence $s$ is assigned a relevance score based on how strongly it presents an opinion about a specific aspect. To account for the fact that a sentence might relate to more than one aspect, [21] calculates this score as an average of its cosine similarity scores (at word level) with all the aspects in $A$. We, however observe from the data that majority of the sentences (especially with strong sentiment polarity) focus on a particular aspect. Further, such averaging reduces the actual strength with which the sentence relates to this aspect. After preprocessing (mainly to discard irrelevant words such as pronouns and stop-words), we obtain the relevance of each word $w_i$ with each of the aspects in $A'$ (by calculating cosine similarity between their embeddings) and finally take the maximum of these values to get the relevance score of the sentence. $Relevance(s)$, Therefore,

$$Relevance(s) = \max_{1 \leq i \leq |s|} \left( \max_{1 \leq j \leq |A'|} \cos(w_i, A_j) \right)$$

We multiply these three scores to obtain the opinion score of the sentence $s$ denoted by $Opin\_Score(s)$. 

---

*Note: The mathematical notations and equations are simplified for clarity. For a detailed understanding, please refer to the original research paper.*
5.2 Opinion Summarization

The goal of our opinion summarization task is to maximize the collective salience of the selected subset of sentences while minimizing redundancy. Additionally, our framework allows readers to customize the summaries according to their needs. Let the desired length of summary be $L$ words. Further, let $fp$ represent the desired percentage of female opinions to be included in the summary. We formalize these objectives as a constrained maximization problem using ILP. In the following equations, $l_i$ gives us the length of sentence $s_i$ in words. $f_i$ and $m_i$ are complements of each other where $f_i = 1$ if $s_i$ is an opinion made by a female reviewer and $m_i = 1$ if it is otherwise. $x_i$ is a binary indicator variable whose value indicates whether to include sentence $s_i$ in the final summary. $sim_{ij}$ represents the cosine similarity between the embeddings of sentences $s_i$ and $s_j$. Similar to [21], equations 5 and 6 ensure that $y_{ij}$, another binary variable, is 1 if and only if both $x_i$ and $x_j$ are 1. We use Gurobi [1] to maximize the objective defined in Equation 2.

$$\text{argmax} \sum_{i} \text{Opin}\_\text{Score}(s_i) x_i - \sum_{ij} \text{sim}_{ij} y_{ij} - C$$

$$\text{subject to: } \sum_{i} l_{i} x_i \leq L$$

$$y_{ij} \leq \frac{1}{2}(x_i + x_j) \ \forall \ i, j$$

Equation 4 restricts the length of the summary to $L$ words. Apart from maximizing the collective opinion scores of the selected sentences, Equation 2 tries to minimize their collective mutual similarity scores. It also minimizes the absolute difference between the desired ratio $fp$ and the actual ratio of female to male comments in the final summary, as defined by Equation 3.

The goal of our opinion summarization task is to maximize the collective salience of the selected subset of sentences while minimizing redundancy. Additionally, our framework allows readers to customize the summaries according to their needs. Let the desired length of summary be $L$ words. Further, let $fp$ represent the desired percentage of female opinions to be included in the summary. We formalize these objectives as a constrained maximization problem using ILP. In the following equations, $l_i$ gives us the length of sentence $s_i$ in words. $f_i$ and $m_i$ are complements of each other where $f_i = 1$ if $s_i$ is an opinion made by a female reviewer and $m_i = 1$ if it is otherwise. $x_i$ is a binary indicator variable whose value indicates whether to include sentence $s_i$ in the final summary. $sim_{ij}$ represents the cosine similarity between the embeddings of sentences $s_i$ and $s_j$. Similar to [21], equations 5 and 6 ensure that $y_{ij}$, another binary variable, is 1 if and only if both $x_i$ and $x_j$ are 1. We use Gurobi [1] to maximize the objective defined in Equation 2.

$$\text{argmax} \sum_{i} \text{Opin}\_\text{Score}(s_i) x_i - \sum_{ij} \text{sim}_{ij} y_{ij} - C$$

$$\text{subject to: } \sum_{i} l_{i} x_i \leq L$$

$$y_{ij} \leq \frac{1}{2}(x_i + x_j) \ \forall \ i, j$$

Equation 4 restricts the length of the summary to $L$ words. Apart from maximizing the collective opinion scores of the selected sentences, Equation 2 tries to minimize their collective mutual similarity scores. It also minimizes the absolute difference between the desired ratio $fp$ and the actual ratio of female to male comments in the final summary, as defined by Equation 3.

6 EXPERIMENTAL SETUP AND RESULTS

Dataset: We create a diverse dataset by collecting all (English) reviews posted on TripAdvisor (till July 30, 2019) for the New Seven Wonders of the World. For each review scraped, we consider the following fields for our analysis: id, text, user rating (1-5), #likes (no. of likes received), username and reviewer’s location. We use geopy 1 to obtain reviewer’s country of origin from his/her location information. We further use the genderComputer 2 package to determine reviewer’s gender from his/her ‘username’ and ‘country of origin’. As majority of usernames on TripAdvisor are not well-formed names, gender could be properly identified for around 20% of the reviews, on average, across the seven places. Among these, we include the top 1000 liked reviews for each place in our final dataset, statistics of which are shown in Table 1.

Experiments on Aspect Identification: We want aspect discovery to be an unsupervised one-time process. In order to obtain aspects/topics which are common across various tourist destinations, we merge the data for all the seven places under consideration and perform our experiments using default parameter settings as mentioned in [10]. We experiment with several values of $K$ ranging from 10 to 50 and obtain the best set of interpretable topics for $K = 25$. We manually label each of the identified topics with fine-grained aspect labels. We finally categorize the obtained topics into coarse-grained aspect classes and find them to be well-aligned with the tourism literature [13]. Result of this many-to-many mapping is reported in Table 2. Since aspect extraction is not a major contribution of this work, we do not evaluate our obtained results against a gold-standard annotated corpus.

Experiments on Summarization: While assigning salience scores to opinions in Section 5.1, we obtain the aspect (class) embeddings as the average of Word2Vec embeddings of the top 10 aspect terms from each topic belonging to that class. For calculating sentence similarity scores, we use Sentence-BERT [15], a state-of-the-art method for obtaining sentence embeddings.

We compare our results with two recent unsupervised and extractive summarization techniques, FairSumm [7] and Centroid [16],3 and with Opinosis [9], an ablative method for opinion summarization. As per standard practices, we employ ROUGE to evaluate the summaries. We note that the original reviews are themselves short textual accounts of user opinions about a place. Therefore, we rank them based on the #likes received, and consider the text from the top 10 liked reviews for each place as a proxy for gold summaries. As none of our baselines facilitate aspect-based summary generation, we compare our default summaries ($L = 100, fp = 0.5$, set of all aspects $A$) and report ROUGE Precision scores in Table 3, macro-averaged across all places. We also conduct ablation studies to explore the effect of various constraints and opinion-scoring modules and report our results in Table 3.

Since these scores are tentative due to unavailability of gold summaries, we perform a human evaluation of our summaries using crowdsourcing. Specifically, we create anonymous survey forms

---

1https://github.com/geopy/geopy
2https://github.com/tue-mdse/genderComputer
3We also considered [3, 21] but their codes/files/instructions are yet to be released.

---

| Place | No. of Reviews |
|-------|----------------|
| Whole data | 3164 (94.3%) |
| The Roman Colosseum (Rome) | 492 (14.9%) |
| Christ the Redeemer (Brazil) | 445 (14.4%) |
| Machu Picchu (Peru) | 461 (14.9%) |
| Petra (Jordan) | 439 (14.4%) |
| Taj Mahal (India) | 498 (16.2%) |
| Chichen Itza (Mexico) | 482 (15.8%) |
| Great Wall of China (China) | 452 (14.2%) |

| Aspect Classes | Inferred Topics |
|----------------|----------------|
| Attractions | Architecture, Monuments, Surroundings, Events |
| Access | Travel infrastructure, Long-distance modes of transport, Entrance, Time of visiting |
| Activities | Photography, Events, Shopping |
| Amenities | Tour Guides, Hospitality, Vendors, Services, Information, Food, Accommodation |
| Culture | Hospitality, Events, Climate, History, Dress |
| Cost | Cost of visiting, Tickets |
| Negatives | Bad Experiences, Vendors |
| Miscellaneous | Nationality, Adjectives |

---

Table 1: Tourist Review Dataset Statistics

Table 2: Mapping between (coarse-grained) aspect classes and (fine-grained) inferred topics
### Table 3: Comparison of ROUGE Scores.

| Methods  | ROUGE-1 | ROUGE-2 | ROUGE-L |
|----------|---------|---------|---------|
| Opinosis | 74.8    | 21.3    | 49.6    |
| Centroid | 62.3    | 14.2    | 38.8    |
| FairSumm | 67.0    | 15.2    | 41.0    |

### Table 4: Results of the crowdsourcing-based evaluation, micro-averaged over all the seven places. Values indicate (%) of times a method is preferred for a particular question.

| Methods  | All Aspects | Access | Amenities, Culture |
|----------|-------------|--------|--------------------|
|          | Q1  | Q2  | Q3  | Q1  | Q2  | Q3  | Q1  | Q2  | Q3  |
| Opinosis | 0.1 | 0.09| 0.07| 0.08| 0.08| 0.08| 0.12| 0.09| 0.08|
| Centroid | 0.14| 0.15| 0.1  | 0.08| 0.1| 0.13| 0.13| 0.09| 0.21|
| FairSumm | 0.13| 0.12| 0.15| 0.23| 0.23| 0.24| 0.13| 0.24| 0.13|
| Our Method | 0.63 | 0.61 | 0.65 | 0.64 | 0.34 | 0.36 | 0.02 | 0.56 | 0.56 |

---

7 CONCLUSION AND FUTURE WORK

To our knowledge, this is the first attempt at producing personalized aspect-based opinion summaries, using an unsupervised extractive summarization framework, which additionally maintains fair representation of opinions from male and female reviewers (default setting). We motivate the need for such summaries and create a tourist review dataset for our experiments. We further establish the effectiveness of our framework by comparing it with recent unsupervised methods for opinion summarization. In future, we would like to experiment with the data of lesser known places.

ACKNOWLEDGMENTS

This research is supported by IMPRINT-2, a national initiative of the Ministry of Human Resource Development (MHRD), India.

REFERENCES

[1] 2015. Gurobi – The overall fastest and best supported solver available. http://www.gurobi.com/
[2] Reinald Kim Amplayo and Mirella Lapata. 2019. Informative and Controllable Opinion Summarization. ArXiv abs/1909.02322 (2019).
[3] Stefanos Angelidis and Mirella Lapata. 2018. Summarizing Opinions: Aspect Extraction Meets Sentiment Prediction and They Are Both Weakly Supervised. In EACL.
[4] Arthur Braizinskis, Marela Lapata, and Ivan Titov. 2019. Unsupervised Multi-Document Opinion Summarization as Copycat-Review Generation. ArXiv abs/1911.02247 (2019).
[5] Eric Chu and Peter J. Liu. 2019. MeanSum: A Neural Model for Unsupervised Multi-Document Abstractive Summarization. In ICMCL.
[6] Roque Enrique Lopez Condore and Thiago Alexandre Salgueiro Pardo. 2017. Opinion summarization methods: Comparing and extending extractive and abstractive approaches. Expert Syst. Appl. 78 (2017), 124–134.
[7] Abhishek Dash, Anurag Shandilya, Arindam Biswas, Kripabandhu Ghosh, Saptarsi Ghosh, and Abhijan Chakraborty. 2019. Summarizing User-generated Textual Content: Motivation and Methods for Fairness in Algorithmic Summaries. Proceedings of the ACM on Human-Computer Interaction 3 (2019), 1–28.
[8] Angela Fan, David Grangier, and Michael Auli. 2017. Controllable Abstractive Summarization. In NMT/IACL.
[9] Kavita Ganesan, Chengxiang Zhai, and Jiawei Han. 2010. Opinionis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions. In COLING.
[10] Ruudin He, Wei Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An Unsupervised Neural Attention Model for Aspect Extraction. In ACL.
[11] Mingjing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In KDD.
[12] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In ACL.
[13] Edison Marrese-Taylor, Juan D. Velásquez, and Felipe Bravo-Marquez. 2014. A novel deterministic approach for aspect-based opinion mining in tourism products reviews. Expert Syst. Appl. 41 (2014), 7764–7775.
[14] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2016. SummaRuNNer: A Recursive Neural Network Based Sequence Model for Extractive Summarization of Documents. In AAAI.
[15] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In EMNLP/IJCNLP.
[16] Gaetano Rossiello, Pierpaolo Basile, and Giovano Semeraro. 2017. Centroid-based Text Summarization through Compositionality of Word Embeddings. In Multilang@EACL.
[17] Koustav Rudra, Pawan Goyal, Niloy Ganguly, Prasenjit Mitra, and Mohammad Imran. 2018. Identifying Sub-events and Summarizing Disaster-Related Information from Microblogs. In SIGIR.
[18] Yufei Tian, Jianfei Yu, and Jing Jiang. 2019. Aspect and Opinion Aware Abstractive Review Summarization with Reinforced Hard Typed Decoder. Proceedings of the 28th ACM International Conference on Information and Knowledge Management (2019).
[19] Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2018. Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction. In ACL.
[20] Victoria Yaneva. 2015. Easy-read Documents as a Gold Standard for Evaluation of Text Simplification Output. In RANLP.
[21] Chao Zhao and Snigdha Chaturvedi. 2020. Weakly-Supervised Opinion Summarization by Leveraging External Information. ArXiv abs/1911.09844 (2020).