Beyond Opinion Mining: Summarizing Opinions of Customer Reviews

Reinald Kim Amplayo[1] Arthur Bražinskas[2] Yoshihiko Suhara[3] Xiaolan Wang[3] Bing Liu[4]
[1]Google Research [2]University of Edinburgh
[3]Megagon Labs [4]University of Illinois at Chicago

ABSTRACT

Customer reviews are vital for making purchasing decisions in the Information Age. Such reviews can be automatically summarized to provide the user with an overview of opinions. In this tutorial, we present various aspects of opinion summarization that are useful for researchers and practitioners. First, we will introduce the task and major challenges. Then, we will present existing opinion summarization solutions, both pre-neural and neural. We will discuss how summarizers can be trained in the unsupervised, few-shot, and supervised regimes. Each regime has roots in different machine learning methods, such as auto-encoding, controllable text generation, and variational inference. Finally, we will discuss resources and evaluation methods and conclude with the future directions. This three-hour tutorial will provide a comprehensive overview over major advances in opinion summarization. The listeners will be well-equipped with the knowledge that is both useful for research and practical applications.

ACM Reference Format:
Reinald Kim Amplayo[1] Arthur Bražinskas[2] Yoshihiko Suhara[3] Xiaolan Wang[3] Bing Liu[4], [1]Google Research [2]University of Edinburgh, [3]Megagon Labs [4]University of Illinois at Chicago . 2022. Beyond Opinion Mining: Summarizing Opinions of Customer Reviews. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (SIGIR '22).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

People in the Information Age read reviews from online review websites when making decisions to buy a product or use a service. The proliferation of such reviews has driven research on opinion mining [21, 39], where the ultimate goal is to glean information from multiple reviews so that users can make decisions more effectively. Opinion mining has assumed several facets in its history: among others, there are sentiment analysis [40], that reduces a single review into a sentiment label, opinion extraction [34], that produces a list of aspect-sentiment pairs representing opinions mentioned in the reviews, and most notably *opinion summarization* [47], which creates a textual summary of opinions that are found in multiple reviews about a certain product or service. Opinion summarization is arguably the most effective solution for opinion mining, especially when assisting the user in making decisions. Specifically, textual opinion summaries provide users with information that is both more concise and more comprehensible compared to other alternatives. Thus, opinion mining research on the IR community has geared its focus towards opinion summarization in recent years (see Table 1).

The task of summarizing opinions in multiple reviews can be divided into two subtasks: opinion retrieval and summary generation. Opinion retrieval selects opinions from the reviews that are salient and thus need to be included in the summary. Summary generation produces a textual summary given the retrieved opinions that is concise yet informative and comprehensible for users to read and make decisions effectively. The summary can be generated from scratch with possibly novel tokens (i.e., *abstractive* summarization; [13, 17]) or spans of text directly extracted from the input (i.e., *extractive* summarization; [6, 20]). Traditionally, these subtasks correspond to a pipeline of natural language generation models [12, 32, 47] where opinion retrieval and summary generation are treated as content selection and surface realization tasks, respectively. Thanks to advancements in neural networks, most of the recent methods use an end-to-end approach [9, 10, 13] where both opinion retrieval and summary generation are done by a single model optimized to produce well-formed and informative summaries.

There are two broad types of challenges in opinion summarization: *annotated data scarcity* and *usability*. As reviews-summary pairs are expensive to create, this has resulted in annotated dataset scarcity. However, the exceptional performance of neural networks for text summarization is mostly driven by large-scale supervised training [41, 50], which makes opinion summarization challenging. The second challenge – usability – stems from a number of practical requirements for industrial applications. First, for real-world products and service we often need to summarize many thousands of reviews. This is largely infeasible due to the high computational and memory costs of modelling that many reviews with neural architectures [7]. Second, state-of-the-art text summarizers are prone to hallucinations [31]. In other words, a summarizer might mistakenly generate a summary with information not covered by input reviews, thus misinforming the user. Third, generic summaries not tailored to specific user needs have lesser value. This calls for ways to learn summarizers producing personalized summaries.

This opens exciting avenues to develop methods for solving these major challenges in opinion summarization. In this light, the aim of the tutorial is to inform interested researchers and practitioners, especially in opinion mining and text summarization, about recent and ongoing efforts to improve the state of the art and make opinion summarization systems useful in real-world scenarios. And the tutorial will make the audience well-equipped for addressing these challenges in terms of methods, ideas, and related work.

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

Table 1: Opinion summarization solutions that will be covered in this tutorial. A dagger † denotes that the solution also leverages weak supervision.

Pre-Neural Solutions
Extractive:
LexRank [16], TextRank [33], MEAD [12], Wang et.al [48]
Abstractive:
Opinosis [17], SEA [12], Gerani et.al [18]
Autoencoders
Extractive:
MATE+MT† [6], Mukherjee et.al† [35], ASPMEM† [51], QT [5]
Abstractive:
MeanSum [13], Coavoux et.al [14], OpinionDigest† [44], RecurSum [24], MultimodalSum [22], COOP [23]
Synthetic Training
Abstractive:
Copycat [10], DenoiseSum [3], MMDS [43], Elsahar et.al† [15], Jiang et.al [25], PlanSum [2], TransSum [46], AceSum [1], ConsistSum [26], LSARS [38]
Low-Resource
Abstractive:
Wang et.al [47], FewSum [9], PASS [37], SelSum [11], CondaSum [4], Wei et.al [49]

2 TUTORIAL CONTENT AND OUTLINE

The tutorial will be 3 hours long and consist of the following five parts, which we describe in detail below.

2.1 Part I: Introduction [30 min]

Opinion summarization [21, 27, 45] focuses on summarizing opinionated text, such as customer reviews, and has been actively studied by researchers from the natural language processing and data mining community for decades. There are two major types of opinion summaries: non-textual summaries, such as aggregated ratings [30], aspect-sentiment tables [45], and opinion clusters [20], and textual summaries, which often consist of a short text. Compared to non-textual summaries, which may confuse users due to their complex formats, textual summaries are considered much more user-friendly [36]. Thus, in recent years, the considerable research interest in opinion summarization has shifted towards textual opinion summaries. In this tutorial, we will also focus on recent solutions for generating textual opinion summaries.

Like single document summary [41, 42], textual opinion summary can also be either extractive or abstractive. However, unlike single document summarization, opinion summarization can rarely rely on gold-standard summaries at training time due to the lack of large-scale training examples in the form of review-summary pairs. Meanwhile, the prohibitively many and redundant input reviews also pose new challenges for the task.

In this part of the tutorial, we will first describe the opinion summarization task, its history, and the major challenges that come with the task. We will then provide a brief overview of existing opinion summarization solutions.

2.2 Part II: Solutions To Data Scarcity [90 min]

In this part of the tutorial, we will present multiple existing opinion summarization models, as also summarized in Table 1. These models attempt to solve the annotated data scarcity problem and are classified into four parts: pre-neural models, autoencoder-based models, models that use synthetic data, and models that leverage low-resource annotated data.

2.2.1 Autoencoders [30/90 min]. Due to the lack of training examples, one major approach is to use autoencoders for unsupervised opinion summarization. The autoencoder model consists of an encoder that transforms the input into latent representations and a decoder that attempts to *reconstruct* the original input using a reconstruction objective. It has a wide range of applications in both computer vision and natural language processing community [8, 19, 28]. Autoencoders can also help models obtain better text representations, which allows easier text clustering, aggregation, and selection. Thus, it benefits both extractive and abstractive solutions. In this tutorial, we will first introduce the basics of autoencoders and then describe how to use autoencoders for both extractive and abstractive opinion summarization.

2.2.2 Synthetic Dataset Creation [30/90 min]. The supervised training of high-capacity models on large datasets containing hundreds of thousands of document-summary pairs is critical to the recent success of deep learning techniques for abstractive summarization [41, 42]. The absence of human-written summaries in a large-scale calls for creative ways to synthesize datasets for supervised training of abstractive summarization models. Customer reviews, available in large quantities, can be used to create synthetic datasets for training. Such datasets are created by sampling one review as a pseudo-summary, and then selecting or generating a subset of reviews as input to be paired with the pseudo-summary. Subsequently, the summarizer is trained in a supervised manner to predict the pseudo-summary given the input reviews. This *self-supervised* approach, as has been shown in a number of works [50, *inter alia*], is effective for training summarizers to generate abstractive opinion summamaaries. In this tutorial, we will introduce various techniques to create synthetic datasets, contrast them, and present results achieved by different works.

2.2.3 Low-Resource Learning [30/90 min]. Modern deep learning methods rely on large amounts of annotated data for training. Unlike synthetic datasets, automatically created from customer reviews, annotated datasets require expensive human effort. Consequently, only datasets with a handful of human-written summaries are available, which lead to a number of few-shot models. These models alleviate annotated data-scarcity using specialized mechanisms, such as parameter subset fine-tuning and summary candidate ranking. An alternative to human-written are editor-written summaries that are scraped from the web and linked to customer reviews. This setup is challenging because each summary can have hundreds of associated reviews. In this tutorial, we will present both methods that are few-shot learners and that scale to hundreds of input reviews.

2.3 Part III: Improving Usability [30 min]

In order to make opinion summarizers more useful in industrial settings, a number of features need to be improved. In this part of

the tutorial, we will discuss the following three major features and recent solutions the community has proposed:

- **Scalability** This refers to the ability to handle a massive number of input reviews. To handle large scale input, the ability to retrieve salient information, e.g., reviews or opinions, becomes a important yet challenging feature for opinion summarization solutions.
- **Input Faithfulness** This refers to the ability of a summarizer to generate summaries covered in content by input reviews. In other words, the summarizer should not confuse entities or introduce novel content into summaries.
- **Controllability** This refers to the ability to produce constrained summaries, such as a hotel summary that only includes room cleanliness or a product summary that only covers the negative opinions.

2.4 Part IV: Evaluation and Resources [20 min]

As is common in other areas of natural language processing, in opinion summarization, researchers often rely on automatic metrics. These metrics, such as ROUGE [29], are based on word overlaps with the reference summary. However, word overlap metrics are limited and can weakly correlate with human judgment. To address these shortcomings, human evaluation is often used, where human annotators assess various aspects of generated summaries. In this tutorial, we will present different kinds of human evaluation experiments, how they are designed, and how they are performed.

2.5 Part V: Future work [10 min]

To conclude the tutorial, we will present several notable open questions for opinion summarization, such as the need for additional annotated resources, common issues with the generated summary (e.g., repetition, hallucination, coherency, and factuality), and the ability to handle various type of input data (e.g., images and knowledge bases). Based on these open questions, we will also present future work on opinion summarization.

3 OBJECTIVES

In this tutorial, we will cover a wide range of techniques from pre-neural approaches to the most recent advances for opinion summarization. In addition, we will also introduce the commonly used resources and evaluation metrics. Our goal for this tutorial is to increase the interest of the IR community towards the opinion summarization problem and help researchers to start working on relevant problems.

4 RELEVANCE TO THE IR COMMUNITY

Sentiment analysis has been a major research area in the IR community. Since the tutorial will cover cutting-edge research in the field, it would attract a wide variety of IR researchers and practitioners. We would also like to emphasize that the interest in opinion mining and summarization techniques in the IR community has been rapidly and significantly increased in recent years. We believe that we are the first to offer a tutorial that covers the series of recent opinion summarization models.[1]

[1]More than 80% of the papers were published within the last three years.

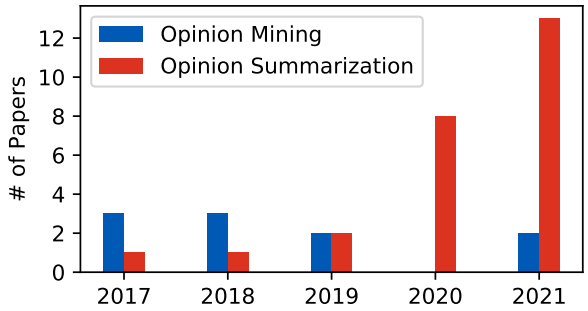

Figure 1: Increasing # of papers for opinion summarization that are published in IR-related venues.

5 SUPPORTING MATERIALS

We will share all materials (e.g., slides and leaderboard website) with the participants before the tutorial, which will be also made publicly available.

6 INSTRUCTORS

The following will present the tutorial:

Reinald Kim Amplayo. is a Research Scientist at Google. He received his PhD from the University of Edinburgh, where his thesis focused on controllable and personalizable opinion summarization. He is a recepient of a best student paper runner-up at ACML 2018.

Arthur Bražinskas. is a PhD student at the University of Edinburgh, supervised by Ivan Titov and Mirella Lapata. He focuses on abstractive opinion summarization using variational methods.

Yoshihiko Suhara. is a Senior Research Scientist at Megagon Labs. He was an Adjunct Instructor at New College of Florida, where he taught a full-semester Deep Learning course for graduate students. He was previously a Visiting Scientist at the MIT Media Lab (2014-2016) and a Research Scientist at NTT Laboratories (2008-2014). He received his PhD from Keio University in 2014. His expertise lies in NLP, especially Opinion Mining and Information Extraction.

Xiaolan Wang. is a Senior Research Scientist at Megagon Labs. She received her PhD from University of Massachusetts Amherst in 2019. Her research interests include data integration, data cleaning, and natural language processing. She co-instructed the tutorial, *Data Augmentation for ML-driven Data Preparation and Integration*, at VLDB 2021.

Bing Liu. is a Distinguished Professor of Computer Science at the University of Illinois at Chicago (UIC). He has published extensively in top conferences and journals. He also authored four books about lifelong learning, sentiment analysis and Web mining. Three of his papers received Test-of-Time awards: two from SIGKDD and one from WSDM. He has served as the Chair of ACM SIGKDD from 2013-2017, as program chair of many leading data mining conferences, including KDD, ICDM, CIKM, WSDM, SDM, and PAKDD, and as associate editor of leading journals such as TKDE, TWEB, DMKD and TKDD. He is a recipient of ACM SIGKDD Innovation Award, and he is a Fellow of the ACM, AAAI, and IEEE.

REFERENCES

[1] Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-Controllable Opinion Summarization. In *EMNLP*. 6578–6593.

[2] Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Unsupervised Opinion Summarization with Content Planning. In *AAAI*, Vol. 35. 12489–12497.

[3] Reinald Kim Amplayo and Mirella Lapata. 2020. Unsupervised Opinion Summarization with Noising and Denoising. In *ACL*. 1934–1945.

[4] Reinald Kim Amplayo and Mirella Lapata. 2021. Informative and Controllable Opinion Summarization. In *EACL*. 2662–2672.

[5] Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *TACL* 9 (2021), 277–293.

[6] Stefanos Angelidis and Mirella Lapata. 2018. Summarizing Opinions: Aspect Extraction Meets Sentiment Prediction and They Are Both Weakly Supervised. In *EMNLP*. 3675–3686.

[7] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv:2004.05150* (2020).

[8] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating Sentences from a Continuous Space. In *CoNLL*. 10–21.

[9] Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Few-Shot Learning for Opinion Summarization. In *EMNLP*. 4119–4135.

[10] Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised Opinion Summarization as Copycat-Review Generation. In *ACL*. 5151–5169.

[11] Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2021. Learning Opinion Summarizers by Selecting Informative Reviews. In *EMNLP*. 9424–9442.

[12] Giuseppe Carenini, Raymond Ng, and Adam Pauls. 2006. Multi-Document Summarization of Evaluative Text. In *EACL*. 305–312.

[13] Eric Chu and Peter Liu. 2019. Meansum: a neural model for unsupervised multi-document abstractive summarization. In *ICML*. PMLR, 1223–1232.

[14] Maximin Coavoux, Hady Elsahar, and Matthias Gallé. 2019. Unsupervised aspect-based multi-document abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. 42–47.

[15] Hady Elsahar, Maximin Coavoux, Jos Rozen, and Matthias Gallé. 2021. Self-Supervised and Controlled Multi-Document Opinion Summarization. In *EACL*. 1646–1662.

[16] Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research* 22 (2004), 457–479.

[17] Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions. In *COLING*. Coling 2010 Organizing Committee, 340–348.

[18] Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *EMNLP*. 1602–1613.

[19] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. 2011. Transforming auto-encoders. In *ICANN*. Springer, 44–51.

[20] Minqing Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews. In *KDD*. 168–177.

[21] Minqing Hu and Bing Liu. 2006. Opinion extraction and summarization on the web. In *AAAI*, Vol. 7. 1621–1624.

[22] Jinbae Im, Moonki Kim, Hoyeop Lee, Hyunsouk Cho, and Sehee Chung. 2021. Self-Supervised Multimodal Opinion Summarization. In *ACL*. 388–403.

[23] Hayate Iso, Xiaolan Wang, Yoshihiko Suhara, Stefanos Angelidis, and Wang-Chiew Tan. 2021. Convex Aggregation for Opinion Summarization. In *EMNLP Findings*. 3885–3903.

[24] Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. 2021. Unsupervised Abstractive Opinion Summarization by Generating Sentences with Tree-Structured Topic Guidance. *TACL* 9 (09 2021), 945–961.

[25] Wenjun Jiang, Jing Chen, Xiaofei Ding, Jie Wu, Jiawei He, and Guojun Wang. 2021. Review Summary Generation in Online Systems: Frameworks for Supervised and Unsupervised Scenarios. *ACM Trans. Web* 15, 3, Article 13 (May 2021), 33 pages.

[26] Wenjun Ke, Jinhua Gao, Huawei Shen, and Xueqi Cheng. 2022. ConsistSum: Unsupervised Opinion Summarization with the Consistency of Aspect, Sentiment and Semantic. In *WSDM*. 467–475.

[27] Hyun Duk Kim, Kavita Ganesan, Parikshit Sondhi, and ChengXiang Zhai. 2011. *Comprehensive Review of Opinion Summarization*. Technical Report. University of Illinois at Urbana-Champaign.

[28] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. *CoRR* abs/1312.6114 (2014).

[29] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, 74–81.

[30] Yue Lu, ChengXiang Zhai, and Neel Sundaresan. 2009. Rated aspect summarization of short comments. In *WWW*. 131–140.

[31] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *ACL*. 1906–1919.

[32] Kathleen McKeown. 1992. *Text generation*. Cambridge University Press.

[33] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *EMNLP*. 404–411.

[34] Arjun Mukherjee and Bing Liu. 2012. Aspect Extraction through Semi-Supervised Modeling. In *ACL*. 339–348.

[35] Rajdeep Mukherjee, Hari Chandana Peruri, Uppada Vishnu, Pawan Goyal, Sourangshu Bhattacharya, and Niloy Ganguly. 2020. Read what you need: Controllable aspect-based opinion summarization of tourist reviews. In *SIGIR*. 1825–1828.

[36] G. Murray, E. Hoque, and G. Carenini. 2017. Chapter 11 - Opinion Summarization and Visualization. In *Sentiment Analysis in Social Networks*, Federico Alberto Pozzi, Elisabetta Fersini, Enza Messina, and Bing Liu (Eds.). Morgan Kaufmann, Boston, 171–187.

[37] Nadav Oved and Ran Levy. 2021. PASS: Perturb-and-Select Summarizer for Product Reviews. In *ACL*. 351–365.

[38] Haojie Pan, Rongqin Yang, Xin Zhou, Rui Wang, Deng Cai, and Xiaozhong Liu. 2020. Large scale abstractive multi-review summarization (lsars) via aspect alignment. In *SIGIR*. 2337–2346.

[39] Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2, 1-2 (2008), 1–135.

[40] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *EMNLP*. 79–86.

[41] Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *EMNLP*. 379–389.

[42] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *ACL*. 1073–1083.

[43] Ori Shapira and Ran Levy. 2020. Massive multi-document summarization of product reviews with weak supervision. *arXiv preprint arXiv:2007.11348* (2020).

[44] Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. 2020. OpinionDigest: A Simple Framework for Opinion Summarization. In *ACL*. 5789–5798.

[45] Ivan Titov and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *ACL:HLT*. 308–316.

[46] Ke Wang and Xiaojun Wan. 2021. TransSum: Translating Aspect and Sentiment Embeddings for Self-Supervised Opinion Summarization. In *ACL Findings*. 729–742.

[47] Lu Wang and Wang Ling. 2016. Neural Network-Based Abstract Generation for Opinions and Arguments. In *NAACL*. 47–57.

[48] Lu Wang, Hema Raghavan, Claire Cardie, and Vittorio Castelli. 2014. Query-Focused Opinion Summarization for User-Generated Content. In *COLING*. 1660–1669.

[49] Penghui Wei, Jiahao Zhao, and Wenji Mao. 2021. A Graph-to-Sequence Learning Framework for Summarizing Opinionated Texts. *TASLP* 29 (2021), 1650–1660.

[50] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *ICML (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 11328–11339.

[51] Chao Zhao and Snigdha Chaturvedi. 2020. Weakly-supervised opinion summarization by leveraging external information. In *AAAI*, Vol. 34. 9644–9651.