Summarizing Contrastive Themes via Hierarchical Non-Parametric Processes

Zhaochun Ren
z.ren@uva.nl

Maarten de Rijke
derijke@uva.nl

University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT

Given a topic of interest, a contrastive theme is a group of opposing pairs of viewpoints. We address the task of summarizing contrastive themes: given a set of opinionated documents, select meaningful sentences to represent contrastive themes present in those documents. Several factors make this a challenging problem: unknown numbers of topics, unknown relationships among topics, and the extraction of comparative sentences. Our approach has three core ingredients: contrastive theme modeling, diverse theme extraction, and contrastive theme summarization. Specifically, we present a hierarchical non-parametric model to describe hierarchical relations among topics; this model is used to infer threads of topics as themes from the nested Chinese restaurant process. We enhance the diversity of themes by using structured determinantal point processes for selecting a set of diverse themes with high quality. Finally, we pair contrastive themes and employ an iterative optimization algorithm to select sentences, explicitly considering contrast, relevance, and diversity. Experiments on three datasets demonstrate the effectiveness of our method.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering

Keywords

Contrastive theme summarization; Structured determinantal point processes; Hierarchical sentiment-LDA; Topic modeling

1. INTRODUCTION

In recent years multi-document summarization has become a well studied task for helping users understanding a set of documents. Typically, the focus has been on relatively long, factual and grammatically correct documents [6, 17, 25, 41, 44, 48]. However, the web now holds a large number of opinionated documents, especially in opinion pieces, microblogs, question answering platforms and web forum threads. The growth in volume of such opinionated documents on the web motivates the development of methods to facilitate the understanding of subjective viewpoints present in sets of documents.

Given a set of opinionated documents, we define a *viewpoint* to be a topic with a specific sentiment label, following [37]. A *theme* is a set of viewpoints around a specific set of topics and an explicit sentiment opinion. Given a set of specific topics, two themes are *contrastive* if they are related to the topics, but opposite in terms of sentiment. The phenomenon of contrastive themes is widespread in opinionated web documents [8]. In Fig. 1 we show an example of three contrastive themes about the "Palestine and Israel relationship." Here, each pair of contrastive themes includes two sentences representing two relevant but opposing themes. In this paper, our focus is on developing methods for automatically detecting and describing contrastive themes.

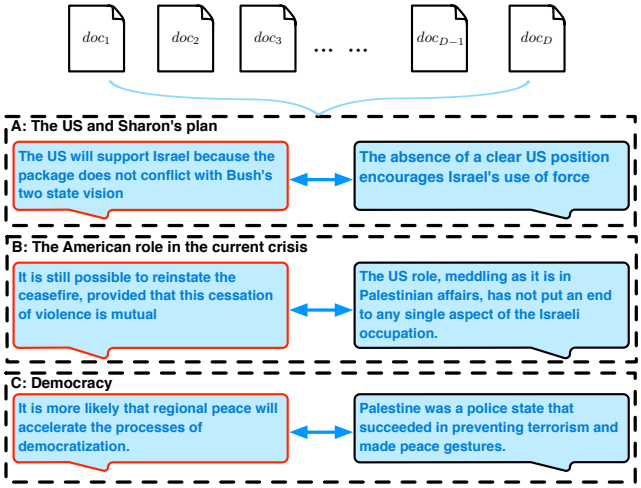

Figure 1: Three example contrastive themes related to "Palestine and Israel." Each contrastive theme shows a pair of opposing sentences.

The task on which we focus is *contrastive summarization* [18, 37] of multiple themes. The task is similar to *opinion summarization*, in which opinionated documents are summarized into structured or semi-structured summaries [12, 13, 15, 19]. However, most existing opinion summarization strategies are not adequate for summarizing contrastive themes from a set of unstructured documents. To our knowledge, the most similar task in the literature is the *contrastive viewpoint summarization* task [37], in which the authors extract contrastive but relevant sentences to reflect contrastive topic aspects, which are derived from a latent topic-aspect model [36]. However, their proposed method for *contrastive viewpoint sum-*

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.

ACM 978-1-4503-3621-5/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2766462.2767713.

marization neglects to explicitly model the number of topics and the relations among topics in contrastive topic modeling—these are two key features in contrastive theme modeling.

The specific contrastive summarization task that we address in this paper is *contrastive theme summarization of multiple opinionated documents*. In our case, the output consists of contrastive sentence pairs that highlight every contrastive theme in the given documents. To address this task, we employ a non-parametric strategy based on the nested Chinese restaurant process (nCRP) [4]. Previous work has proved the effectiveness of non-parametric models in topic modeling [1, 39]. But none of them considers the task of contrastive theme summarization. We introduce a topic model that aims to extract contrastive themes and describe hierarchical relations among the underlying topics. Each document in our model is represented by hierarchical threads of topics, whereas a word in each document is assigned a finite mixture of topic paths. We apply collapsed Gibbs sampling to infer approximate posterior distributions of themes.

To enhance the diversity of the contrastive theme modeling, we then proceed as follows. Structured determinantal point processes (SDPPs) [21] are a novel probabilistic strategy to extract diverse and salient threads from large data collections. Given theme distributions obtained via hierarchical sentiment topic modeling, we employ SDPPs to extract a set of diverse and salient themes. Finally, based on themes extracted in the first two steps, we develop an iterative optimization algorithm to generate the final contrastive theme summary. During this process, *relevance*, *diversity* and *contrast* are considered.

Our experimental results, obtained using three publicly available opinionated document datasets, show that contrastive themes can be successfully extracted from a given corpus of opinionated documents. Our proposed method for multiple contrastive themes summarization outperforms state-of-the-art baselines, as measured using ROUGE metrics.

To sum up, our contributions in this paper are as follows:

- We focus on a contrastive theme summarization task to summarize contrastive themes from a set of opinionated documents.
- We apply a hierarchical non-parametric model to extract contrastive themes for opinionated texts. We tackle the diversification challenge by employing structured determinantal point processes to sample diverse themes.
- Jointly considering relevance, diversity and contrast, we apply an iterative optimization strategy to summarize contrastive themes, which is shown to be effective in our experiments.

We introduce related work in §2. We formulate our research problem in §3 and describe our approach in §4. Then, §5 details our experimental setup and §6 presents the experimental results. Finally, §7 concludes the paper.

2. RELATED WORK

2.1 Multi-document summarization

Multi-document summarization (MDS) is useful since it is able to provide a brief digest of large numbers of relevant documents on the same topic [34]. Most existing work on MDS is based on the extractive format, where the target is to extract salient sentences to construct a summary. Both unsupervised and supervised based learning strategies have received lots of attention. One of the most widely used unsupervised strategies is clustering with respect to the centroid of the sentences within a given set of documents; this idea has been applied by NeATS [28] and MEAD [38]. Many other recent publications on MDS employ graph-based ranking methods [10]. Wan and Yang [48] propose a theme-cluster strategy based on conditional Markov random walks. Similar methods are also applied in [49] for a query-based MDS task. Celikyilmaz and Hakkani-Tur [6] consider the summarization task as a supervised prediction problem based on a two-step hybrid generative model, whereas the Pythy summarization system [47] learns a log-linear sentence ranking model by combining a set of semantic features. As to discriminative models, CRF-based algorithms [44] and structured SVM-based classifiers [25] have proved to be effective in extractive document summarization. Learning to rank models have also been employed to query-based MDS [43] and to topic-focused MDS [50]. In recent years, with the development of social media, multi-document summarization is being applied to social documents, e.g., tweets, weibos, and Facebook posts [7, 9, 35, 40, 41]. Temporal and update summarization [2] is becoming a popular task in MDS research [34]; for this task one follows a stream of documents over time and summarizes information on what is new compared to what has been summarized previously [31, 35, 45].

2.2 Opinion summarization

In recent years, *opinion* summarization has received extensive attention. Opinion summarization generates structured [15, 24, 30, 32] or semi-structured summaries [13, 16, 20] given opinionated documents as input. Opinosis [12] generates a summary from redundant data sources. Similarly, a graph-based multi-sentence compression approach has been proposed in [11]. Meng et al. [33] propose an entity-centric topic-based opinion summarization framework, which is aimed at generating summaries with respect to topics and opinions.

Other relevant work for our contrastive summarization has been published by Lerman and McDonald [23] and Paul et al. [37]. Lerman and McDonald [23] propose an approach to extract representative contrastive descriptions from product reviews. A joint model between sentiment mining and topic modeling is applied in [37].

2.3 Non-parametric topic modeling

Non-parametric topic models are aimed at handling infinitely many topics; they have received much attention. For instance, hierarchical Latent Dirichlet Allocation (hLDA) [4] describes latent topics over nested Chinese restaurant processes. To capture the relationship between latent topics, nested Chinese restaurant processes generate tree-like topical structures over documents. Traditional non-parametric topic models do not explicitly address diversification among latent variables during clustering. To tackle this issue, Kulesza and Taskar [21, 22] propose a stochastic process named structured determinantal point process (SDPP), where diversity is explicitly considered. As an application in text mining, Gillenwater et al. [14] propose a method for topic modeling based on SDPPs. As far as we know, the determinantal point process has not been integrated with other non-parametric models yet.

To the best of our knowledge, there is little previous work on summarizing contrastive themes. In this paper, by optimizing the number of topics, building relations among topics and enhancing the diversity among themes, we propose a hierarchical topic modeling strategy to summarize contrastive themes in the given documents.

3. PRELIMINARIES

3.1 Problem formulation

Before introducing our method for contrastive theme summarization, we introduce our notation and key concepts. Table 1 lists the notation we use in this paper.

Table 1: Glossary.

Symbol	Description
$\mathcal{D}$	candidate documents
$\mathcal{W}$	vocabulary in corpus $\mathcal{D}$
$\mathcal{K}$	themes set in $\mathcal{D}$
$\mathcal{T}$	themes tuples from $\mathcal{K}$
d	a document, $d \in \mathcal{D}$
s_d	a sentence in document d, i.e., $s_d \in d$
w	a word present in a sentence, $w \in \mathcal{W}$
x	a sentiment label, $x \in \{neg, neu, pos\}$
o_s	sentiment distribution of sentence s
c^x	a topic path under label x
b	a topic node on a topic path
z^x	a topic level under x label
ϕ^x	topic distribution of words, under label x
$k_{c,x}$	a theme corresponding to topic path c, under label x
t	a contrastive theme tuple
θ_d	probability distribution of topic levels over d
$\mathcal{S}_t$	contrastive summary for theme tuple t

Given a corpus $\mathcal{D}$, we begin by defining the notions of topic, sentiment and theme in our work. Following topic modeling customs [3], we define a *topic* in a document d to be a probability distribution over words. Unlike "flat" topic models [3], we assume that each document d can be represented by multiple topics that are organized in an infinite tree-like hierarchy $c = \{(z_0, c), (z_1, c), \ldots\}$, $z_0 \prec z_1 \prec \ldots$, i.e., c indicates a path from the root topic level z_0 on the infinite tree to more specialized topics that appear at the leaves of the tree, and for each topic level z we define a topic node $b = (z, c)$ on the topic path c.

Sentiment is defined as a probability distribution over sentiment labels *positive*, *negative*, and *neutral*. A *sentiment label* x is attached with each word w. Considering the *sentiment*, we divide topics into three classes: positive topics (2), neutral topics (1) and negative topics (0).

Given all hierarchical topics and sentiment labels, we define a *theme* $k_{c,x}$ as a threaded topic path c from the root level to the leaf level for the given sentiment label x. Let $\mathcal{K}$ be the set of themes, and let $\mathcal{K}^{pos}$, $\mathcal{K}^{neg}$, $\mathcal{K}^{neu}$ indicate the set of positive, negative and neutral themes, respectively, i.e., $\mathcal{K} = \mathcal{K}^{pos} \cup \mathcal{K}^{neg} \cup \mathcal{K}^{neu}$. Furthermore, we define a *contrastive theme* to be a theme tuple $t = (c^{pos}, c^{neg}, c^{neu})$ by extracting themes fromis contained in $\mathcal{K}^{pos} \times \mathcal{K}^{neg} \times \mathcal{K}^{neu}$. Themes c^{pos}, c^{neg} and c^{neu} in each tuple t are relevant in topic but opposite in sentiment labels.

Finally, we define contrastive theme summarization. Given a set of documents $\mathcal{D} = \{d_1, d_2, ..., d_D\}$, the purpose of the contrastive theme summarization task (CTS) is to select a set of meaningful sentences $\mathcal{S}_t = \{S_{c^{pos}}, S_{c^{neg}}, S_{c^{neu}}\}$ to reflect the representative information in each possible theme tuple $t = (c^{pos}, c^{neg}, c^{neu})$.

3.2 Determinantal point process

A point process $\mathcal{P}$ on a discrete set $\mathcal{Y} = \{y_1, y_2, \ldots, y_N\}$ is a probability measure on the power set $2^{\mathcal{Y}}$ of $\mathcal{Y}$. We follow the definitions from [21]. A determinantal point process (DPP) $\mathcal{P}$ is a point process with a positive semidefinite matrix M indexed by the elements of $\mathcal{Y}$, such that if $\mathcal{Y} \sim \mathcal{P}$, then for each $\mathcal{A} \subseteq \mathcal{Y}$, there is $\mathcal{P}(\mathcal{A} \subseteq \mathcal{Y}) = \det(M_{\mathcal{A}})$. Here, $M_{\mathcal{A}} = [M_{i,j}]_{y_i, y_j \in \mathcal{A}}$ is the restriction of M to the entries indexed by elements of $\mathcal{A}$. Matrix M is defined as the marginal kernel, where it contains all information to compute the probability of $\mathcal{A} \subseteq \mathcal{Y}$. For the purpose of modeling data, the construction of DPP is via *L-ensemble* [5]. Using L-ensemble, we have

$$\mathcal{P}(\mathcal{Y}) = \frac{\det(L_{\mathcal{Y}})}{\sum_{\mathcal{Y}' \subset \mathcal{Y}} \det(L_{\mathcal{Y}'})} = \frac{\det(L_{\mathcal{Y}})}{\det(L + I)}, \tag{1}$$

where I is the $N \times N$ identity matrix, L is a positive semidefinite matrix; $L_{\mathcal{Y}} = [L_{i,j}]_{y_i, y_j \in \mathcal{Y}}$ refers to the restriction of L to the entries indexed by elements of $\mathcal{Y}$, and $\det(L_{\emptyset}) = 1$. For each entry of L, we have

$$L_{ij} = q(y_i)\varphi(y_i)^{\mathrm{T}}\varphi(y_j)q(y_j), \tag{2}$$

where $q(y_i) \in \mathbb{R}^+$ is considered as the "quality" of an item y_i; $\varphi(y_i)^{\mathrm{T}}\varphi(y_j) \in [-1, 1]$ measures the similarity between item y_i and y_j. Here, for each $\varphi(y_i)$ we set $\varphi(y_i) \in \mathbb{R}^D$ as a normalized D-dimensional feature vector, i.e., $\|\varphi(y_i)\|_2 = 1$. Because the value of a determinant of vectors is equivalent to the volume of the polyhedron spanned by those vectors, $\mathcal{P}(\mathcal{Y})$ is proportional to the volumes spanned by $q(y_i)\varphi(y_i)$. Thus, sets with high-quality, diverse items will get the highest probability in DPP.

Building on the DPP, *structured* determinantal point processes (SDPPs) have been proposed to efficiently handle the problem containing exponentially many structures [14, 21, 22]. In the setting of SDPPs, items set $\mathcal{Y}$ contains a set of threads of length T. Thus in SDPPs, each item y_i has the form $y_i = \{y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(T)}\}$, where $y_i^{(t)}$ indicates the document at the t-th position of thread y_i. To make the normalization and sampling efficient, SDPPs assume a factorization of $q(y_i)$ and $\varphi(y_i)^{\mathrm{T}}\varphi(y_j)$ into parts, decomposing quality multiplicatively and similarity additively, as follows:

$$q(y_i) = \prod_{t=1}^{T} q(y_i^{(t)}); \quad \varphi(y_i) = \sum_{t=1}^{T} \varphi(y_i^{(t)}); \tag{3}$$

the quality function $q(y_i)$ has a simple log-linear model setting $q(y_i) = \exp(\lambda w(y_i))$, where λ is set as a hyperparameter that balances between quality and diversity. An efficient sampling algorithm for SDPPs has been proposed by Kulesza and Taskar [21].

Since SDPPs specifically address "diversification" and "saliency," we apply them to identify diversified and salient themes from themes sets $\mathcal{K}$. We will detail this step in §4.

4. METHOD

4.1 Overview

We provide a general overview of our method for performing contrastive theme summarization (CTS) in Fig. 2. There are three main phases: (A) contrastive theme modeling; (B) diverse theme extraction; and (C) contrastive theme summarization. To summarize, we are given a set of documents $\mathcal{D} = \{d_1, d_2, \ldots, d_D\}$ as input. For each document $d \in \mathcal{D}$, in phase (A) (see §4.2), we obtain a structured themes set $\mathcal{K}$ with a root node r, topic distributions ϕ and opinion distributions o_s.

In (B) (see §4.3), given the structured output themes $\mathcal{K}$, we employ a structured determinantal point process to obtain a subset $\mathcal{K}' \subseteq \mathcal{K}$ to enhance the saliency and diversity among themes.

Based on themes $\mathcal{K}'$ and their corresponding topic distributions and opinion distributions, in (C) (see §4.4) we generate the final contrastive theme summary $\mathcal{S}$. We develop an iterative optimization algorithm for this process: the first part in §4.4 is to generate the contrastive theme tuples $\mathcal{T}$, each of which includes relevant themes for a topic but contrastive in sentiment; the second part in §4.4 is meant to generate the final contrastive summary $\mathcal{S} = \{S_t\}$ for each theme tuple.

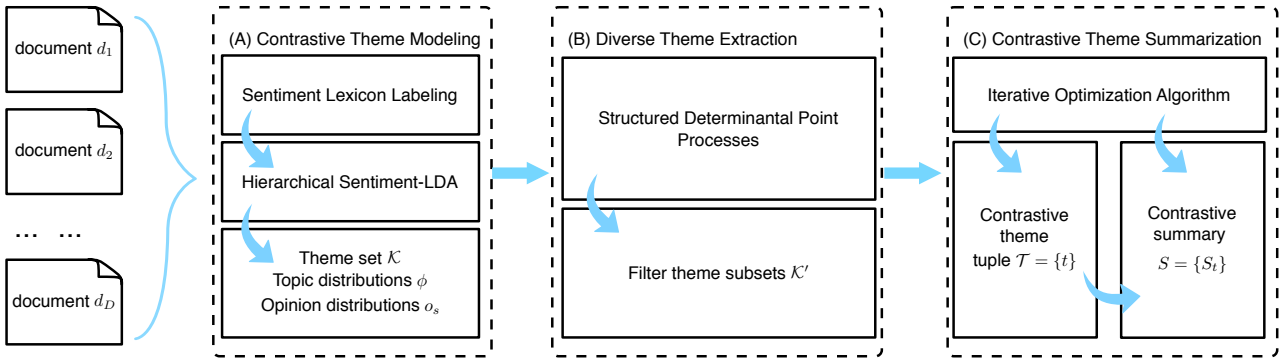

Figure 2: Overview of our approach to contrastive theme summarization. (A) indicates contrastive theme modeling; (B) indicates a structured determinantal point process to diversify topics; and (C) refers to the contrastive summary generation algorithm. Crooked arrows indicate the output in each step; while straight arrows indicate processing directions.

4.2 (A) Contrastive theme modeling

We start by proposing a hierarchical sentiment-LDA model to jointly extract topics and opinions from our input corpus. Unlike previous work on traditional "flat" topic models [37], our method can adaptively generate topics organized in a tree-like hierarchy.

Briefly, each document $d \in \mathcal{D}$ can be represented as a collection of sentences, whereas each sentence $s \in d$ is composed of a collection of words. By using a state-of-the-art sentiment analysis method [46], for each word w in each document d we extract its sentiment label x_w, where $x_w \in \{pos, neu, neg\}$. Generally, for document d we select three threaded topic paths $\{c^x\}$, with $x = pos$, neu, neg, each of which is generated by a nested Chinese restaurant process (nCRP) [4]. After deriving the sentiment label x, each word $w \in d$ is assigned to a specific topic level z by traversing from the root to the leave on the path c^x.

Next, we give a more detailed technical account of our model. Following the nested Chinese restaurant process [4], our topic model identifies documents with threaded topic paths generated by nCRP. Given level z, we consider each node (z, c) on a threaded topic path c as a specific topic. To select the exact topic level $z \in [1, L]$, we draw a variable θ_d from a Dirichlet distribution derived from hyperparameter m, to define a probability distribution on topic levels along the topic path c. Given a draw from a Dirichlet distribution, document d is generated by repeatedly selecting a topic level. We assume that each document $d \in \mathcal{D}$ is represented by three classes of topics: positive, negative and neutral topics.

In document d, for each sentence $s \in d$ we define a sentiment distribution o_s from a Dirichlet distribution over a hyper parameter γ. For each word $w \in \mathcal{W}$, we select three topic levels z^{pos}, z^{neg} and z^{neu} from a discrete distribution over θ_d, respectively. While the sentiment label is derived from a multinomial distribution over o_s, w is derived from a discrete distribution over topic levels $\{z^{pos}, z^{neg}, z^{neu}\}$. The generation process of our proposed model is shown in Fig. 3.

Since exact posterior inference in hierarchical sentiment-LDA is intractable, we employ a collapsed Gibbs sampler to approximate the posterior distributions of topic level z_w for each word w and topic path c_d for each document d. In our model, two sets of variables are observed: the sentiment labels x_w for each word w, and the words set $\mathcal{W}$. Our sampling procedure is divided into two steps for each iteration: (1) sampling a topic path for each document; (2) sampling level allocation for each word.

For the sampling procedure of thread c_d, given current other variables on document d, we have:

$$p(c_d^x | c_{-d}^x, z, o) \propto p(c_d^x | c_{-d}^x) \cdot p(W_d | W_{-d}, c, x, o, z) \quad (4)$$

1. For each topic level $z^x \in \mathcal{Z}^x$ in infinite tree:
   - Draw $\phi^x \sim Dirichlet(\beta^x)$;
2. For each document $d \in \mathcal{D}$:
   - Draw $c_d^x \sim nCRP(p)$;
   - Draw $\theta_d \sim Dirichlet(m)$;
   - For each sentence $s \in d$:
     - Draw opinion $o_s \sim Dirichlet(\gamma)$;
     - For each word $w \in N$:
       - Draw sentiment $x \sim Multinomial(o_s)$;
       - Draw topics $z^x \sim Discrete(\theta_d)$:
       - Draw word $w \sim Discrete(\phi_{z^x, c_d^x})$;

Figure 3: Generative process in hierarchical sentiment-LDA.

where $p(c_d^x | c_{-d}^x)$ in (4) is the prior distribution implied by the nested Chinese restaurant process, whereas for each topic node (z, c_d) on path c_d, we have:

$$\begin{cases} P((z, c_d) = b_i) = \frac{n_i}{n+p-1} \\ P((z, c_d) = b_{new}) = \frac{p}{n+p-1} \end{cases} \quad (5)$$

where b_i indicates a node that has been taken before, b_{new} indicates a new node that has not been considered yet; n_i refers to the number of times that topic node (z, c_d) is assigned to a document. To infer $p(W_d | W_{-d}, c, x, o, z)$, we integrate over multinomial parameters and have:

$$p(W_d | W_{-d}, c, x, o, z) \propto \prod_{z=1}^{L} \left(\frac{\Gamma(n_{-d}^{z,c} + W\beta)}{\prod_{w \in W} \Gamma(n_{w,-d}^{z,c} + \beta)} \frac{\prod_{w \in W} \Gamma(n_{w,-d}^{z,c} + n_{w,d}^{z,c} + \beta)}{\Gamma(n_{-d}^{z,c} + n_d^{z,c} + W\beta)} \prod_{s \in S_d} \frac{\prod_{x \in X} \Gamma(n_{s,x} + \gamma_x)}{\Gamma(n_s + \gamma)} \right), \quad (6)$$

where $n_{-d}^{z,c}$ indicates the number of times that documents have been assigned to topic node (z, c) leaving out document d; $n_{w,-d}^{z,c}$ denotes the number of times that word w has been assigned to the topic node (z, c) leaving out document d.

To sample topic level $z_{d,n}$ for each word w_n in document d, we find its joint probabilistic distribution of terms, sentiment labels and

topics as follows:

$$p(z^x_{d,n} = \eta | z^x_{-(d,n)}, c^x, x, o, w) \propto \frac{n^{\eta,c}_{w_n,-n} + \beta}{n^{\eta,c}_{-n} + W\beta} \frac{n^{\eta}_{d} + m}{n^{\eta}_{d,-n} + Lm} \frac{\prod_{x \in X} \Gamma(n_{s,x} + \gamma_x)}{\Gamma(n_s + \gamma)} \quad (7)$$

where $z^x_{-(d,n)}$ denotes the vectors of level allocations leaving out $z^x_{d,n}$ in document d. Further, $n^{\eta,c}_{w_n,-n}$ denotes the number of times that words have been assigned to topic node (η, c) that are the same as word w_n; $n^{\eta}_{d,-n}$ denotes the number of times that document d have been assigned to level k leaving out word w_n.

After Gibbs sampling, we get a set of topic paths $\{c^x\}$ that can be represented as themes $\mathcal{K} = \{k_{c,x}\}$; for each word w in d, we have hybrid parametric distributions ϕ^x that reflect the topic distribution given a specific level z on path c, i.e., $P(w, x|c, z) = \phi^x_{z,c,w}$. For each sentence s, we have a probability distribution o_s over sentiment labels, i.e., $P(x|s) = o_{s,x}$.

4.3 (B) Diverse theme extraction

Given a set of themes $\mathcal{K} = \{k_{c,x}\}$ resulting from step (A), some further issues need to be tackled before we arrive at our desired summary. On the one hand, many themes in $\mathcal{K}$ share common topics; on the other hand, many words' topic probabilities ϕ are similar, which makes it difficult to distinguish the importance of the themes.

To address this dual problem, we employ the structured determinantal point process (SDPP) [22] to select a subset of salient and diverse themes from $\mathcal{K}$. Following [21], we define a structured determinantal point process $\mathcal{P}$ as a type of probability distribution over a subset of themes belonging to $\mathcal{K}$. Two main factors are considered in SDPPs: the *quality* q_i and the *similarity* ${\varphi_i}^T \varphi_j$. A subset with high quality and highly diverse themes will be assigned the highest probability $\mathcal{P}$ by the SDPPs.

Given themes $\mathcal{K}$ sampled from (A), we proceed as follows. Firstly, for each theme $k \in \mathcal{K}$ we use $q((z_i, c))$ to indicate the "quality" of topic $(z_i, c) \in k$ and we use $\varphi((z_i, c))^T \varphi((z_j, c')) \in [0, 1]$ to refer to a measure of similarity between two topics (z_i, c) and (z_j, c'):

$$q((z_i, c)) = \sum_{w \in \mathcal{W}_H} \phi_{z_i,c,w};$$
$$\varphi((z_i, c))^T \varphi((z_j, c')) = \exp\left(-\frac{\left\|\Phi_{z_i,c} - \Phi_{z_j,c'}\right\|_2^2}{2\sigma^2}\right), \quad (8)$$

where $\Phi_{z_i,c}$ indicates the vector $\{\phi_{z,c,w}\}_{w \in \mathcal{W}}$; $\left\|\Phi_{z_i,c} - \Phi_{z_j,c'}\right\|_2^2$ is the squared Euclidean distance between $\Phi_{z_i,c}$ and $\Phi_{z_j,c'}$; $\mathcal{W}_H$ indicates the top-n salient words; σ is a free parameter. Based on (1) and (2), we construct the semidefinite matrix $\mathcal{M}$ for SDPPs.

For two topic paths $c_i = \{(z_1, c_i), \ldots, (z_L, c_i)\}$ and $c_j = \{(z'_1, c_j), \ldots, (z_L, c_j)\}$, $c_i, c_j \in \mathcal{K}$, we assume a factorization of the quality $q(c)$ and similarity score $\varphi(c_i, c_j)$ into parts, decomposing quality multiplicatively and similarity additively, i.e., for topic paths c_i and c_j, $q(c_i)$ and $\varphi(c_i, c_j)$ are calculated by (3), respectively.

To infer the posterior results of SDPPs over themes, we adapt an efficient sampling algorithm as described in Algorithm 1. Following [21], we let $\mathcal{M} = \sum_{k=1}^{K} \lambda_k v_k v_k^T$ be an orthonormal eigendecomposition, and let e_i be the ith standard basis K-vector. The sampling algorithm of SDPPs outputs a subset of themes, i.e., $\mathcal{K}' = \{k'_{c,x}\}$, which reflect a trade-off between high quality and high diversity.

Algorithm 1: Sampling process for SDPPs

Input : Eigenvector/values pairs $\{(v_k, \lambda_k)\}$; Themes set $\mathcal{K}$;
Output: Filtered themes set $\mathcal{K}'$ from SDPPs;
$\mathcal{J} \leftarrow 0$; $\mathcal{K}' \leftarrow 0$;
for $k \in \mathcal{K}$ **do**
  $\mathcal{J} \leftarrow \mathcal{J} \cup \{k\}$ with probability $\frac{\lambda_k}{1+\lambda_k}$;
end
$V \leftarrow \{v_k\}_{k \in \mathcal{J}}$;
while $|V| > 0$ **do**
  Select k_i from $\mathcal{K}$ with $P(k_i) = \frac{1}{|V|} \sum_{v \in V} (v^T e_i)^2$;
  $\mathcal{K}' \leftarrow \mathcal{K}' \cup k_i$;
  $V \leftarrow V_\perp$ as an orthonormal basis for the subspace of V orthonormal to e_i;
end
return $\mathcal{K}'$.

4.4 (C) Contrastive theme summarization

In this section, we specify the sentence selection procedure for contrastive themes. Considering the diversity among topics, we only consider leaf topics in each theme $k'_{c,x} \in \mathcal{K}'$. Thus, each theme $k'_{c,x}$ can be represented by a leaf topic (z^x_L, c^x) exclusively. For simplicity, we abbreviate leaf topics sets $\{(z^x_L, c^x)\}$ as $\{c^x\}$.

Given $\{c^x\}$, we need to connect topics in various classes to a set of contrastive theme tuples of the form $t = (c^{pos}_i, c^{neg}_{ii}, c^{(neu)}_{iii})$. To assess the correlation between two topics (c^x_i) and (c^y_{ii}) in different classes, we define a *correlation* based on topic distributions $\Phi_{z,c}$ as follows:

$$1 - \frac{1}{N} \sum_{d \in \mathcal{D}} \left| \sum_{w \in d} \phi_{z_L, c^x_i, w} - \sum_{w' \in d} \phi_{z_L, c^y_{ii}, w'} \right|. \quad (9)$$

We sample three leaf topics from the three classes mentioned earlier (positive, negative and neutral), so that the total *correlation* values for all three topic pairs has maximal values.

Next, we extract representative sentences for each contrastive theme tuple $t = (c^{pos}_i, c^{neu}_{ii}, c^{neg}_{iii})$. An intuitive way for generating the contrastive theme summary is to extract the most salient sentences as a summary. However, high-degree topical relevance cannot be taken as the only criterion for sentence selection. To extract a contrastive theme summary $\mathcal{S}_t = \{S_{c^{pos}_i}, S_{c^{neu}_{ii}}, S_{c^{neg}_{iii}}\}$ for tuple $t = (c^{pos}_i, c^{neu}_{ii}, c^{neg}_{iii})$, in addition to *relevance* we consider two more key requirements *contrast* and *diversity*. Given selected sentences $\mathcal{S}'_t$, we define a salient score $F(s_i|\mathcal{S}'_c, t)$:

$$F(s_i|\mathcal{S}'_t, t) = ctr(s_i|\mathcal{S}'_t, t) + div(s_i, \mathcal{S}'_t) + rel(s_i|t) \quad (10)$$

where $ctr(s_i|\mathcal{S}'_t, t)$ indicates the contrast between s_i and $\mathcal{S}'_t$ for t; $div(s_i, \mathcal{S}'_t)$ indicates the divergence between s_i and $\mathcal{S}'_t$; $rel(s_i|t)$ indicates the relevance of s_i given t.

Contrast calculates the sentiment divergence between the currently selected sentence s_i and the results of extracted sentences set $\mathcal{S}'_t$, under the given theme t. Our intention is to make the current sentence as contrastive as possible from extracted sentences as much as possible. Therefore, we have:

$$ctr(s_i|\mathcal{S}'_t, t) = \max_{\{s \in \mathcal{S}'_t, x\}} \left| (o_{s_i,x} - o_{s,x}) \cdot (\phi^x_{z_L,c,w} - \phi^x_{z_L,c,w}) \right|. \quad (11)$$

Diversity calculates the information divergence among all sentences within the current candidate result set. Ideally, the contrastive summary results have the largest possible difference in theme distribu-

Algorithm 2: Iterative process for generating the summary $\mathcal{S}$.

Input : $\mathcal{T} = \{(c_i^{pos}, c_{ii}^{neg}, c_{iii}^{neu})\}$, μ,π,S,N;
Output: $\mathcal{S} = \left\{\{S_{c_i^{pos}}, S_{c_{ii}^{neg}}, S_{c_{iii}^{neu}}\}_{(t)}\right\}$;

for *each* $t = (c_i^{pos}, c_{ii}^{neg}, c_{iii}^{neu})$ **do**
  Rank and extract relevant sentences to $\mathcal{C}$ by $rel(s|t)$;
  Initialize: Extract $\frac{N}{|\mathcal{T}|}$ sentences from $\mathcal{C}$ to $\mathcal{S}_t$;
  repeat
    Extract $\mathcal{X} = \{s_x \in \mathcal{C} \cap s_x \notin \mathcal{S}_t\}$;
    for $s_x \in \mathcal{X}, \forall s_y \in \mathcal{S}_t$ **do**
      Calculate $\mathcal{L} = \sum_{s_i \in \mathcal{S}_t} F(s_i|\mathcal{S}_t, t)$;
      Calculate
      $\Delta\mathcal{L}_{s_x,s_y} = \mathcal{L}((\mathcal{S}_t - s_y) \cup s_x) - \mathcal{L}(\mathcal{S}_t)$;
    end
    Get $\langle \hat{s_x}, \hat{s_y} \rangle$ that $\langle \hat{s_x}, \hat{s_y} \rangle = \arg\max_{s_x,s_y} \Delta\mathcal{L}_{s_x,s_y}$;
    $\mathcal{S}_t = (\mathcal{S}_t - \hat{s_y}) \cup \hat{s_x}$;
  until $\forall \Delta S_{s_x,s_y} < \varepsilon$;
  $\mathcal{S} = \mathcal{S} \cup \mathcal{S}_t$;
end
return $\mathcal{S}$.

tions with each other. The equation is as follows:

$$div(s_i|\mathcal{S}'_t) = \max_{s \in \mathcal{S}'_t} |rel(s_i|t) - rel(s|t)| \,. \tag{12}$$

Furthermore, a contrastive summary should contain relevant sentences for each theme t, and minimize the information loss with the set of all candidate sentences. Thus, given $\phi^x_{z_L,c,w}$, the *relevance* of sentence s_i given theme t is calculated as follows:

$$rel(s_i|t) = \frac{1}{N_{s_i}} \sum_x \sum_{w \in s_i} \phi^x_{z_L,c,w}. \tag{13}$$

Algorithm 2 shows the details of our sentence extraction procedure.

5. EXPERIMENTAL SETUP

5.1 Research questions

We list the research questions **RQ1**–**RQ4** that guide the remainder of the paper.

RQ1 Is hierarchical sentiment-LDA effective for extracting contrastive themes from documents? (See §6.1.) Is hierarchical sentiment-LDA helpful for optimizing the number of topics during contrastive theme modeling? (See §6.2.)

RQ2 Is the structured determinantal point process helpful for compressing the themes into a diverse and salient subset of themes? (See §6.2 and §6.3.) What is the effect of SDPP in contrastive theme modeling? (See §6.3).

RQ3 How does our iterative optimization algorithm perform on contrastive theme summarization? Does it outperform baselines? (See §6.4.)

RQ4 What is the effect of *contrast*, *diversity* and *relevance* for contrastive theme summarization in our method? (See §6.5.)

5.2 Datasets

We employ three datasets in our experiments. Two of them have been used in previous work [36, 37], and another one is extracted from news articles of the New York Times.[1] All documents in our datasets are written in English. All three datasets include human-made summaries, which are considered as ground-truth in our experiments. As an example, Table 2 shows statistics of 15 themes from the three datasets that include the largest number of articles in our dataset. In total, 15,736 articles are used in our experiments.

Table 2: Top 15 topics in our three datasets. Column 1 shows the name of topic; column 2 shows the number of articles included in the topic; column 3 shows the publication period of those articles, and column 4 indicates to which dataset the topic belongs.

General description	# articles	Period	Dataset
U.S. International Relations	3121	2004–2007	3
Terrorism	2709	2004–2007	3
Presidential Election of 2004	1686	2004	3
U.S. Healthcare Bill	940	2010	1
Budgets & Budgeting	852	2004–2007	3
Israel-Palestine conflict	594	2001–2005	2
Airlines & Airplanes	540	2004–2007	3
Colleges and Universities	490	2004–2007	3
Freedom and Human Rights	442	2004–2007	3
Children and Youth	424	2004–2007	3
Computers and the Internet	395	2004–2007	3
Atomic Weapons	362	2004–2005	3
Books and Literature	274	2004–2007	3
Abortion	170	2004–2007	3
Biological and Chemical Warfare	152	2004–2006	3

The first dataset ("dataset 1" in Table 2) consists of documents from a Gallup[2] phone survey about the 2010 U.S. healthcare bill. It contains 948 verbatim responses, collected March 4–7, 2010. Respondents indicate if they are "for" or "against" the bill, and there is a roughly even mix of the two opinions (45% for and 48% against). Each document in this dataset only includes 1–2 sentences.

Our second dataset ("dataset 2") is extracted from the Bitterlemons corpus, which is a collection of 594 opinionated articles about the Israel-Palestine conflict. The Bitterlemons corpus consists of the articles published on the Bitterlemons website[3] from late 2001 to early 2005. This dataset has also been applied in previous work [29, 36]. Unlike the first dataset, this dataset contains long opinionated articles with well-formed sentences. It too contains a fairly even mixture of two different perspectives: 312 articles from Israeli authors and 282 articles from Palestinian authors.

Our third dataset ("dataset 3") is a set of articles from the New York Times. The New York Times Corpus contains over 1.8 million articles written and published between January 1, 1987 and June 19, 2007. Over 650,000 articles have manually written article summaries. In our experiments, we only use *Opinion* column articles that were published during 2004–2007.

5.3 Baselines and comparisons

We list the methods and baselines that we consider in Table 3. We write HSDPP for the overall process as described in Section 4, which includes steps (A) contrastive theme modeling, (B) diverse theme extraction and (C) contrastive theme summarization. We write HSLDA for the model that only considers steps (A) and (C), so skipping the structured determinantal point processes in (B). To evaluate the effect of *contrast*, *relevance* and *diversity*, we consider HSDPPC, the method that only considers *contrast* in contrastive

[1] http://ilps.science.uva.nl/resources/nyt_cts

[2] http://www.gallup.com/home.aspx

[3] http://www.bitterlemons.org

Table 3: Our methods and baselines used for comparison.

Acronym	Gloss	Reference
HSDPPC	HSDPP only considering *contrast* in (C) contrastive theme summarization	This paper
HSDPPR	HSDPP only considering *relevance* in (C) contrastive theme summarization	This paper
HSDPPD	HSDPP only considering *diversity* in (C) contrastive theme summarization	This paper
HSLDA	Contrastive theme summarization method in (C) with HSLDA, without SDPPs	This paper
HSDPP	Contrastive theme summarization method in (C) with HSLDA and SDPPs	This paper
Topic models		
TAM	Topic-aspect model based contrastive summarization	[36]
Sen-TM	Sentiment LDA based contrastive summarization	[24]
LDA	LDA based document summarization	[3]
HLDA	Hierarchical LDA based document summarization	[4]
Summarization		
LexRank	LexRank algorithm for summarization	[10]
DFS	Depth-first search for sentence extraction	[13]
ClusterCMRW	Clustering-based sentence ranking strategy	[48]

theme summarization. We write HSDPPR for the method that only considers *relevance* and HSDPPD for the method that only considers *diversity* in the summarization.

To assess the contribution of our proposed methods, our baselines include recent related work. For contrastive theme modeling, we use the Topic-aspect model (TAM, [36]) and the Sentiment-topic model (Sen-TM, [24]) as baselines for topic models. Both focus on the joint process between topics and opinions. Other topic models, such as Latent dirichlet allocation (LDA) [3] and hierarchical latent dirichlet allocation (HLDA) [4], are also considered in our experiments. For the above "flat" topic models, we evaluate their performance using varying numbers of topics (10, 30 and 50 respectively). The number of topics used will be shown as a suffix to the model's name, e.g., TAM-10.

We also consider previous document summarization work as baselines: (1) A depth-first search strategy (DFS, [13]) based on our topic model. (2) The LexRank algorithm [10] that ranks sentences via a Markov random walk strategy. (3) ClusterCMRW [48] that ranks sentences via a clustering-based method. (4) Random, which extracts sentences randomly.

5.4 Experimental setup

Following existing models, we set pre-defined values for some parameters in our proposed method. In our proposed hierarchical sentiment-LDA model, we set m as 0.1 and γ as 0.33 as default values in our experiments.

Optimizing the number of topics is a problem shared between all topic modeling approaches. In our hierarchical sentiment-LDA model, we set the default length of L to 10, and we discuss it in our experiments. As same as other non-parametric topic models, our HSLDA model optimizes the number of themes automatically. Under the default settings in our topic modeling, we find that for the Gallup investigation data, the optimal number of topics is 23; the Bitterlemons corpus, it is 67; for the New York Times dataset, it is 282.

5.5 Evaluation metrics

To assess the saliency of contrastive theme modeling in our experiments, we adapt the *purity* and *accuracy* in our experiments to measure performance. To evaluate the diversity among topics we calculate the *diversity* as follows:

$$diversity = \frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} \max \left| \phi^{x}_{z,c,w} - \phi^{x}_{z',c',w} \right| \quad (14)$$

We adopt the ROUGE evaluation metrics [27], a widely-used recall-oriented metric for document summarization that evaluates the overlap between a gold standard and candidate selections. We use ROUGE-1 (R-1, *unigram based method*), ROUGE-2 (R-2, *bigram based method*) and ROUGE-W (R-W, *weighted longest common sequence*) in our experiments.

Statistical significance of observed differences between the performance of two runs is tested using a two-tailed paired t-test and is denoted using $^{\blacktriangle}$ (or $^{\blacktriangledown}$) for strong significance for $\alpha = 0.01$; or $^{\triangle}$ (or $^{\triangledown}$) for weak significance for $\alpha = 0.05$. In our experiments, significant difference are with regard to TAM and TAM-Lex for contrastive theme modeling and contrastive theme summarization, respectively.

6. RESULTS AND DISCUSSION

6.1 Contrastive theme modeling

We start by addressing **RQ1** and test whether HSLDA and HSDPP are effective for the contrastive theme modeling task. First, Table 4 shows an example topic path of our hierarchical sentiment-LDA model. Column 1 shows the topic levels, columns 2, 3 and 4 show the 7 most representative words with positive, neutral and negative sentiment labels, respectively. For each sentiment label, we find semantic dependencies between adjacent levels.

Table 5 compares the *accuracy* and *purity* of our proposed methods to four baselines. We find that HSDPP and HSLDA tend to outperform the baselines. For the *Bitterlemons* and *New York Times* corpora, HSDPP exhibits the best performance both in terms of *accuracy* and *purity*. Compared to TAM, HSDPP shows a 9.5% increase in terms of *accuracy*. TAM achieves the best performance on the *Healthcare Corpus* when we set its number of topics to 10. However, the performance differences between HSDPP and TAM on this corpus are not statistically significant. This shows that our proposed contrastive topic modeling strategy is effective in contrastive topic extraction.

6.2 Number of themes

To start, for research question **RQ1**, to evaluate the effect of the length of each topic path to the performance of contrastive theme modeling, we examine the performance of HSDPP with different values of topic level L, in terms of *accuracy*. In Figure 4, we find that the performance of HSDPP in terms of *accuracy* peaks when the length of L equals 12; with fewer than 12, performance keeps increasing but if the number exceeds 12, due to the redundancy of topics in contrastive summarization, performance decreases.

Unlike TAM and Sen-LDA, HSDPP and HSLDA determine the optimal number of topics automatically. In Table 5 we find that the results for TAM change with various number of topics. However, for HSDPP we find that it remains competitive for all three corpora while automatically determining the number of topics.

6.3 Effect of structured determinantal point processes

Turning to **RQ2**, Table 5 shows that performance of HSDPP and HSLDA on contrastive theme modeling in terms of *accuracy* and *purity*, for all three datasets. We find that HSDPP outperforms HSLDA in terms of both *accuracy* and *purity*. Table 5 also contrasts the evaluation results for HSDPP with TAM and Sen-TM in

Table 4: Part of an example topic path of hierarchical sentiment-LDA result about "College and University." Columns 2, 3 and 4 list popular positive, neutral and negative terms for each topic level, respectively.

Topic level	Positive	Neutral	Negative
1	favor, agree, accept, character paid, interest, encourage	college, university, university school, editor, year	lost, suffer, fish, wrong, ignore drawn, negative
2	education, grant, financial, benefit save, recent, lend, group	Harvard, president, summer, Lawrence university, faculty, term, elite	foreign, hard, low global, trouble lose, difficulty
3	attract, meaningful, eligible, proud essence, quarrel,qualify	summers, Boston, greek, season seamlessly, opinion, donation	short, pity, unaware, disprove disappoint, idiocy, disaster
4	practical, essay, prospect respect, piously, behoove	write, march, paragraph, analogy analogy, Princeton, english	dark, huge, hassle, poverty depression, inaction, catastrophe
5	grievance, democratic, dignity, elite interest, frippery, youthful	June, volunteer, community, Texas classmate, liberal, egger	cumbersome, inhumane, idiocy, cry mug, humble, hysteria

Table 5: RQ1 and RQ2: *Accuracy*, *purity* and *diversity* values for contrastive theme modeling. Significant differences are with respect to TAM-10 (row with shaded background).

	Healthcare Corpus			Bitterlemons Corpus			New York Times		
	accuracy	*purity*	*diversity*	*accuracy*	*purity*	*diversity*	*accuracy*	*purity*	*diversity*
LDA-10	0.336$^{\blacktriangledown}$	0.337$^{\blacktriangledown}$	0.156$^{\triangledown}$	0.346$^{\blacktriangledown}$	0.350$^{\blacktriangledown}$	0.167$^{\triangledown}$	0.321$^{\blacktriangledown}$	0.322$^{\blacktriangledown}$	0.172$^{\blacktriangledown}$
LDA-30	0.313$^{\blacktriangledown}$	0.315$^{\blacktriangledown}$	0.134$^{\blacktriangledown}$	0.324$^{\blacktriangledown}$	0.332$^{\blacktriangledown}$	0.137$^{\blacktriangledown}$	0.317$^{\blacktriangledown}$	0.317$^{\blacktriangledown}$	0.144$^{\blacktriangledown}$
LDA-50	0.294$^{\blacktriangledown}$	0.298$^{\blacktriangledown}$	0.115$^{\blacktriangledown}$	0.304$^{\blacktriangledown}$	0.309$^{\blacktriangledown}$	0.121$^{\blacktriangledown}$	0.295$^{\blacktriangledown}$	0.301$^{\blacktriangledown}$	0.134$^{\blacktriangledown}$
TAM-10	**0.605**	0.602	0.222	0.645	0.646	0.241	0.551	0.560	0.271
TAM-30	0.532$^{\triangledown}$	0.534$^{\triangledown}$	0.194	0.623	0.626	0.224	0.564	0.564	0.242
TAM-50	0.522$^{\triangledown}$	0.525$^{\triangledown}$	0.152	0.596$^{\triangledown}$	0.596$^{\triangledown}$	0.174	0.576	0.582	0.195$^{\blacktriangledown}$
Sen-TM-10	0.530$^{\triangledown}$	0.531	0.194	0.537$^{\blacktriangledown}$	0.539$^{\blacktriangledown}$	0.209	0.514	0.518	0.255
Sen-TM-30	0.484$^{\blacktriangledown}$	0.488$^{\blacktriangledown}$	0.184	0.492$^{\blacktriangledown}$	0.502$^{\blacktriangledown}$	0.163$^{\triangledown}$	0.473	0.478	0.195$^{\blacktriangledown}$
Sen-TM-50	0.471$^{\blacktriangledown}$	0.481$^{\blacktriangledown}$	0.164	0.479$^{\blacktriangledown}$	0.482$^{\blacktriangledown}$	0.152$^{\triangledown}$	0.454$^{\blacktriangledown}$	0.456$^{\blacktriangledown}$	0.182$^{\blacktriangledown}$
HLDA	0.324$^{\blacktriangledown}$	0.326$^{\blacktriangledown}$	0.223	0.346$^{\blacktriangledown}$	0.342$^{\blacktriangledown}$	0.263	0.329$^{\blacktriangledown}$	0.330$^{\blacktriangledown}$	0.291
HSLDA	0.591	0.598	0.225	0.658	0.660	0.269	0.573	0.578	0.292
HSDPP	0.603	**0.604**	**0.244**	**0.692**	**0.696**	**0.292**$^{\blacktriangle}$	**0.609**$^{\blacktriangle}$	**0.610**$^{\blacktriangle}$	**0.326**$^{\blacktriangle}$

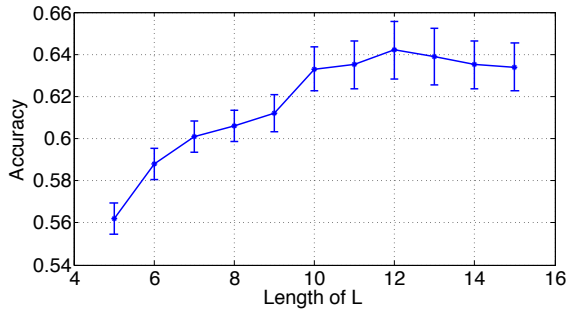

Figure 4: RQ1: Performance with different values of hierarchical topic level *L*, in terms of accuracy

terms of diversity (columns 4, 7, 10). We evaluate the performance of TAM and Sen-TM by varying the number of topics. HSDPP achieves the highest diversity scores. The diversity scores for TAM and Sen-TM decrease as the number of topics increases. In Table 6, we see that HSDPP outperforms HSLDA for all top 15 topics in our dataset in terms of diversity. In terms of diversity, HSDPP offers a significant increase over HSLDA of up to 18.2%.

To evaluate the performance before and after structured determinantal point processes in terms of *accuracy*, Table 6 contrasts the evaluation results for HSDPP with those of HSLDA, which excludes structured determinantal point processes, in terms of *accuracy*. We find that HSDPP outperforms HSLDA for each topic listed in Table 6. In terms of *accuracy*, HSDPP offers a significant increase over HSLDA of up to 14.6%. Overall, HSDPP outperforms HSLDA with a 5.6% increase in terms of accuracy. Hence, we conclude that the structured determinantal point processes helps to enhance the performance of contrastive theme extraction.

6.4 Overall performance

To help us answer **RQ3**, Table 7 lists the ROUGE performance for all summarization methods. As expected, Random performs worst. Using a depth-first search-based summary method (DFS) does not perform well in our experiments. Our proposed method HSDPP significantly outperforms the baselines on two datasets, whereas on the *healthcare corpus* the LexRank-based method performs better than HSDPP, but not significantly. A manual inspection of the outcomes indicates that the contrastive summarizer in HSDPP (i.e., step (C) in Fig. 2) is being outperformed by the LexRank summarizer in HSDPP-Lex on the *Healthcare* dataset because of the small vocabulary and the relative shortness of the documents in this dataset (at most two sentences per document). The summarizer in HSDPP prefers longer documents and a larger vocabulary. We can see this phenomenon on the *Bitterlemons Corpus*, which has 20–40 sentences per document, where HSDPP achieves a 10.3% (13.4%) increase over TAM-Lex in terms of ROUGE-1 (ROUGE-2), whereas the ROUGE-1 (ROUGE-2) score increases 2.2% (4.8%) over HSDPP-Lex. On the *New York Times*, HSDPP offers a significant improvement over TAM-Lex of up to 13.2% and 18.2% in terms of ROUGE-1 and ROUGE-2, respectively.

6.5 Contrastive summarization

Several factors play a role in our proposed summarization method, HSDPP. To determine the contribution of *contrast*, *relevance* and *diversity*, Table 8 shows the performance of HSDPPD, HSDPPR,

Table 7: RQ3: ROUGE performance of all approaches to contrastive document summarization. Significant differences are with respect to TAM-Lex (row with shaded background).

	Healthcare Corpus			Bitterlemons Corpus			New York Times		
	ROUGE-1	ROUGE-2	ROUGE-W	ROUGE-1	ROUGE-2	ROUGE-W	ROUGE-1	ROUGE-2	ROUGE-W
Random	0.132$^{\blacktriangledown}$	0.022$^{\blacktriangledown}$	0.045$^{\blacktriangledown}$	0.105$^{\blacktriangledown}$	0.019$^{\blacktriangledown}$	0.038$^{\blacktriangledown}$	0.102$^{\blacktriangledown}$	0.015$^{\blacktriangledown}$	0.033$^{\blacktriangledown}$
ClusterCMRW	0.292$^{\blacktriangledown}$	0.071$^{\blacktriangledown}$	0.155$^{\blacktriangledown}$	0.263$^{\blacktriangledown}$	0.065$^{\blacktriangledown}$	0.106	0.252$^{\blacktriangledown}$	0.066$^{\blacktriangledown}$	0.098$^{\blacktriangledown}$
DSF	0.264$^{\blacktriangledown}$	0.064$^{\blacktriangledown}$	0.125$^{\blacktriangledown}$	0.235$^{\blacktriangledown}$	0.054$^{\blacktriangledown}$	0.091$^{\blacktriangledown}$	0.211$^{\blacktriangledown}$	0.047$^{\blacktriangledown}$	0.088$^{\blacktriangledown}$
Sen-TM-Lex	0.312$^{\blacktriangledown}$	0.077$^{\triangledown}$	0.141	0.296$^{\blacktriangledown}$	0.062$^{\blacktriangledown}$	0.129	0.284$^{\blacktriangledown}$	0.057$^{\blacktriangledown}$	0.122
TAM-Lex	0.397	0.085	**0.147**	0.362	0.071	0.135	0.341	0.068	0.125
HSDPP	**0.398**	**0.089**	0.142	**0.404**$^{\blacktriangle}$	**0.082**$^{\blacktriangle}$	**0.159**$^{\blacktriangle}$	**0.393**$^{\blacktriangle}$	**0.082**$^{\blacktriangle}$	**0.149**$^{\blacktriangle}$

Table 6: RQ2: Effect of structured determinantal point processes in topic modeling for the top 15 topics in our datasets. Acc. abbreviates accuracy, Div. abbreviates diversity.

	HSLDA		HSDPP	
Descriptions	Acc.	Div.	Acc.	Div.
U.S. Inter. Relations	0.532	0.294	**0.583**$^{\triangle}$	**0.312**
Terrorism	0.569	0.301	**0.621**$^{\blacktriangle}$	**0.341**$^{\blacktriangle}$
2004 Election	0.591	0.266	**0.641**$^{\blacktriangle}$	**0.281**
US. Healthcare	0.591	0.225	**0.603**	**0.244**
Budget	0.506	0.248	**0.551**$^{\blacktriangle}$	**0.299**$^{\triangle}$
Israel-Palestine	**0.658**	0.269	0.652	**0.292**
Airlines	0.602	0.325	0.602	**0.384**$^{\blacktriangle}$
Universities	**0.596**	0.207	0.562	**0.219**
Human Rights	0.571	0.199	**0.624**$^{\triangle}$	**0.206**$^{\triangle}$
Children	**0.712**	0.352	0.622	**0.394**$^{\blacktriangle}$
Internet	0.547	0.277	**0.601**$^{\blacktriangle}$	**0.298**
Atomic Weapons	0.614	0.292	**0.662**$^{\triangle}$	**0.306**$^{\triangle}$
Literature	0.555	0.212	**0.611**$^{\triangle}$	**0.255**$^{\triangle}$
Abortions	0.594	0.301	**0.608**	**0.322**$^{\triangle}$
Bio.&Chemi. warfare	0.596	0.275	**0.597**	**0.302**$^{\triangle}$
Overall	0.581	0.296	**0.614**$^{\triangle}$	**0.317**$^{\triangle}$

and HSDPPC in terms of the ROUGE metrics. We find that HSDPP, which combines *contrast*, *relevance* and *diversity*, outperforms the other approaches on all corpora. After HSDPP, HSDPPR, which includes *relevance* during the summarization process, performs best. Thus, from Table 8 we conclude that *relevance* is the most important part during the summarization process.

7. CONCLUSION

We have considered the task of contrastive theme summarization of multiple opinionated documents. We have identified two main challenges: unknown number of topics and unknown relationships among topics. We have tackled these challenges by combining the nested Chinese restaurant process with contrastive theme modeling, which outputs a set of threaded topic paths as themes. To enhance the diversity of contrastive theme modeling, we have presented the structured determinantal point process to extract a subset of diverse and salient themes. Based on the probabilistic distributions of themes, we generate contrastive summaries subject to three key criteria: contrast, diversity and relevance. In our experiments, we have demonstrated the effectiveness of our proposed method, finding significant improvements over state-of-the-art baselines tested with three manually annotated datasets. Contrastive theme modeling is helpful for extracting contrastive themes and optimizing the number of topics. We have also shown that structured determinantal point processes are effective for diverse theme extraction.

Although we focused mostly on news articles or news-relate articles, our methods are more broadly applicable to other settings with opinionated and conflicted content, such as comment sites or product reviews. Limitations of our work include its ignorance of word dependencies and, being based on hierarchical LDA, the documents that our methods work with should be sufficiently large.

As to future work, parallel processing methods may enhance the efficiency of our topic model on large-scale opinionated documents. Also, the transfer of our approach to streaming corpora should give new insights. It is interesting to consider recent studies such as [26] on search result diversification for selecting salient and diverse themes. Finally, supervised and semi-supervised learning can be used to improve the accuracy in contrastive theme summarization [42].

Acknowledgments. This research was supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nr 312827 (VOX-Pol), the Netherlands Organisation for Scientific Research (NWO) under project nrs 727.011.005, 612.001.116, HOR-11-10, 640.006.013, 612.066.930, CI-14-25, SH-322-15, Amsterdam Data Science, the Dutch national program COMMIT, the ESF Research Network Program ELIAS, the Elite Network Shifts project funded by the Royal Dutch Academy of Sciences (KNAW), the Netherlands eScience Center under project number 027.012.105, the Yahoo! Faculty Research and Engagement Program, the Microsoft Research PhD program, and the HPC Fund.

8. REFERENCES

[1] A. Ahmed, L. Hong, and A. Smola. Nested chinese restaurant franchise process: Applications to user tracking and document modeling. In *ICML*, 2013.

[2] J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In *SIGIR*, 2003.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Machine Learning research*, 3:993–1022, 2003.

[4] D. M. Blei, T. L. Griffiths, and M. I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2), 2010.

[5] A. Borodin. Determinantal point processes. In *The Oxford Handbook of Random Matrix Theory*. Oxford University Press, 2009.

[6] A. Celikyilmaz and D. Hakkani-Tur. A hybrid hierarchical model for multi-document summarization. In *ACL*, 2010.

[7] D. Chakrabarti and K. Punera. Event summarization using tweets. In *ICWSM*, 2011.

[8] S. Dori-Hacohen and J. Allan. Detecting controversy on the web. In *CIKM*, 2013.

[9] Y. Duan, F. Wei, C. Zhumin, Z. Ming, and Y. Shum. Twitter topic summarization by ranking tweets using social influence and content quality. In *Coling*, 2012.

[10] G. Erkan and D. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artificial Intelligence Research*, 22:457–479, 2004.

Table 8: RQ4: ROUGE performance of all our proposed methods in contrastive document summarization. Significant differences are with respect to the row labeled HSDPPD, with shaded background.

	Healthcare Corpus			Bitterlemons Corpus			New York Times		
	ROUGE-1	ROUGE-2	ROUGE-W	ROUGE-1	ROUGE-2	ROUGE-W	ROUGE-1	ROUGE-2	ROUGE-W
HSDPPD	0.291	0.054	0.133	0.301	0.045	0.136	0.284	0.042	0.132
HSDPPR	0.392$^{\blacktriangle}$	0.082$^{\blacktriangle}$	0.138	0.394$^{\blacktriangle}$	0.079$^{\blacktriangle}$	0.146$^{\blacktriangle}$	0.376$^{\blacktriangle}$	0.072$^{\blacktriangle}$	0.147$^{\blacktriangle}$
HSDPPC	0.362	0.078	0.136	0.319	0.059	0.136	0.308	0.067	0.141
HSDPP	**0.398**$^{\blacktriangle}$	**0.089**$^{\blacktriangle}$	**0.142**$^{\triangle}$	**0.404**$^{\blacktriangle}$	**0.082**$^{\blacktriangle}$	**0.159**$^{\blacktriangle}$	**0.393**$^{\blacktriangle}$	**0.082**$^{\blacktriangle}$	**0.149**$^{\blacktriangle}$

[11] K. Filippova. Multi-sentence compression: Finding shortest paths in word graphs. In *Coling*, 2010.

[12] K. Ganesan, C. Zhai, and J. Han. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Coling*, 2010.

[13] K. Ganesan, C. Zhai, and E. Viegas. Micropinion generation: an unsupervised approach to generating ultra-concise summaries of opinions. In *WWW*, 2012.

[14] J. Gillenwater, A. Kulesza, and B. Taskar. Discovering diverse and salient threads in document collections. In *EMNLP-CoNLL*, 2012.

[15] M. Hu and B. Liu. Mining opinion features in customer reviews. In *AAAI*, 2004.

[16] M. Hu and B. Liu. Opinion extraction and summarization on the web. In *AAAI*, 2006.

[17] X. Huang, X. Wan, and J. Xiao. Comparative news summarization using concept-based optimization. *Knowledge and Information Systems*, 38(3):691–716, 2013.

[18] H. D. Kim and C. Zhai. Generating comparative summaries of contradictory opinions in text. In *CIKM*, 2009.

[19] H. D. Kim, K. Ganesan, P. Sondhi, and C. Zhai. Comprehensive review of opinion summarization. *Tech. Rep*, 2011.

[20] H. D. Kim, M. G. Castellanos, M. Hsu, C. Zhai, U. Dayal, and R. Ghosh. Ranking explanatory sentences for opinion summarization. In *SIGIR*, 2013.

[21] A. Kulesza and B. Taskar. Structured determinantal point processes. In *NIPS*, 2010.

[22] A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *Found. & Tr. Machine Learning*, 5(2–3): 123–286, 2012.

[23] K. Lerman and R. McDonald. Contrastive summarization: an experiment with consumer reviews. In *NAACL*, 2009.

[24] F. Li, C. Han, M. Huang, X. Zhu, Y.-J. Xia, S. Zhang, and H. Yu. Structure-aware review mining and summarization. In *Coling*, 2010.

[25] L. Li, K. Zhou, G. Xue, H. Zha, and Y. Yu. Enhancing diversity, coverage and balance for summarization through structure learning. In *WWW*, 2009.

[26] S. Liang, Z. Ren, and M. de Rijke. Fusion helps diversification. In *SIGIR*, 2014.

[27] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *ACL*, 2004.

[28] C.-Y. Lin and E. Hovy. From single to multi-document summarization: A prototype system and its evaluation. In *ACL*, 2002.

[29] W.-H. Lin, T. Wilson, J. Wiebe, and A. Hauptmann. Which side are you on? Identifying perspectives at the document and sentence levels. In *CoNLL*, 2006.

[30] Y. Lu, C. Zhai, and N. Sundaresan. Rated aspect summarization of short comments. In *WWW*, 2009.

[31] R. McCreadie, C. Macdonald, and I. Ounis. Incremental update summarization: Adaptive sentence selection based on prevalence and novelty. In *CIKM*, 2014.

[32] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *WWW*, 2007.

[33] X. Meng, F. Wei, X. Liu, M. Zhou, S. Li, and H. Wang. Entity-centric topic-oriented opinion summarization in twitter. In *KDD*, 2012.

[34] A. Nenkova and K. McKeown. Automatic summarization. *Found. & Tr. Information Retrieval*, 5(2-3):103–233, 2012.

[35] J. Nichols, J. Mahmud, and C. Drews. Summarizing sporting events using twitter. In *IUI*, 2012.

[36] M. Paul and R. Girju. A two-dimensional topic-aspect model for discovering multi-faceted topics. *AAAI*, 2010.

[37] M. J. Paul, C. Zhai, and R. Girju. Summarizing contrastive viewpoints in opinionated text. In *EMNLP*, 2010.

[38] D. Radev et al. MEAD–A platform for multidocument multilingual text summarization. In *LREC*, 2004.

[39] L. Ren, D. B. Dunson, and L. Carin. The dynamic hierarchical dirichlet process. In *ICML*, 2008.

[40] Z. Ren, J. Ma, S. Wang, and Y. Liu. Summarizing web forum threads based on a latent topic propagation process. In *CIKM*, 2011.

[41] Z. Ren, S. Liang, E. Meij, and M. de Rijke. Personalized time-aware tweets summarization. In *SIGIR*, 2013.

[42] Z. Ren, M.-H. Peetz, S. Liang, W. van Dolen, and M. de Rijke. Hierarchical multi-label classification of social text streams. In *SIGIR*, 2014.

[43] C. Shen and T. Li. Learning to rank for query-focused multi-document summarization. In *ICDM*, 2011.

[44] D. Shen, J. Sun, H. Li, Q. Yang, and Z. Chen. Document summarization using conditional random fields. In *IJCAI*, 2007.

[45] L. Shou, Z. Wang, K. Chen, and G. Chen. Sumblr: continuous summarization of evolving tweet streams. In *SIGIR*, 2013.

[46] R. Socher et al. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, 2013.

[47] K. Toutanova, C. Brockett, M. Gamon, J. Jagarlamudi, H. Suzuki, and L. Vanderwende. The pythy summarization system: Microsoft research at DUC 2007. In *DUC*, 2007.

[48] X. Wan and J. Yang. Multi-document summarization using cluster-based link analysis. In *SIGIR*, 2008.

[49] F. Wei, W. Li, Q. Lu, and Y. He. Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization. In *SIGIR*, 2008.

[50] Y. Zhu, Y. Lan, J. Guo, P. Du, and X. Cheng. A novel relational learning-to-rank approach for topic-focused multi-document summarization. In *ICDM*, 2013.