Front.Comput.Sci.
DOI

REVIEW ARTICLE

A Unified Latent Variable Model for Contrastive Opinion Mining

Ebuka IBEKE, Chenghua LIN(✉), Adam WYNER, Mohamad Hardyman BARAWI

Department of Computing Science
University of Aberdeen, UK

Abstract There are large and growing textual corpora in which people express contrastive opinions about the same topic. This has led to an increasing number of studies about contrastive opinion mining. However, there are several notable issues with the existing studies. They mostly focus on mining contrastive opinions from multiple data collections, which need to be separated into their respective collections beforehand. In addition, existing models are opaque in terms of the relationship between topics that are extracted and the sentences in the corpus which express the topics; this opacity does not help us understand the opinions expressed in the corpus. Finally, contrastive opinion is mostly analysed qualitatively rather than quantitatively. This paper addresses these matters and proposes a novel unified latent variable model (contraLDA), which: mines contrastive opinions from both single and multiple data collections, extracts the sentences that project the contrastive opinion, and measures the strength of opinion contrastiveness towards the extracted topics. Experimental results show the effectiveness of our model in mining contrasted opinions, which outperformed our baselines in extracting coherent and informative sentiment-bearing topics. We further show the accuracy of our model in classifying topics and sentiments of textual data, and we compared our results to five strong baselines.

Keywords Contrastive opinion mining, Sentiment analysis, Topic modelling

Received month dd, yyyy; accepted month dd, yyyy

E-mail: *chenghua.lin@abdn.ac.uk* (Corresponding author)

1 Introduction

Recent years have seen a growing interest in text mining applications aimed at uncovering public opinions and social trends. This is partially driven by the fact that the Web now holds a large number of opinionated documents, such as opinion pieces and product reviews, to name a few. An additional driver is that the language one uses to express opinion indicates one's subjective viewpoints; this language can be used to understand and cluster people's opinion based on belief, experience or emotion, rather than facts. Text mining methods are therefore desired for facilitating automatic discovery of subjective viewpoints present in such large amounts of opinionated documents.

We define contrastive opinion mining as the discovery of opinion perspectives held by different individuals or groups, which are related to a given topic but opposite in terms of sentiments. The usefulness of contrastive opinion mining spans across many applications such as discovering the public's stand on major socio-political events [1], observing heated debates over controversial issues where different sides defend their viewpoints with contrasting statements [2], as well as mining issues from product review sites that can serve as an important source of feedback to businesses [3]. For example, there were heated discussions on the web about whether one should install the Mac OS X El Capitan soon after it was released to the public. Table 1 shows some discussions from the Apple Store, where people express highly controversial opinions after upgrading to the system, i.e., some experienced

Table 1 Contrastive opinions regarding El Capitan upgrade.

+ Opinion	Best OS X in terms of speed since Snow Leopard. Fast on start up and no delays. Apps open with lightening speed.
- Opinion	And it is also slow compare to Yosemite IMO. It slowed down my Macbook Pro significantly. Unbelievably slow, runs like garbage now.

pleasant performance improvements while others witnessed a significant drop in speed. Considering the huge number of reviews available, it is highly desirable to acquire an overview of the major viewpoints from large amounts of text data automatically, allowing one to convert data into actionable knowledge and then make decisions in a timely manner.

Recently, mining contrastive opinions has been applied to a variety of tasks, including analysing editorial differences between multiple media sources [4], extracting contrastive viewpoints from political debates [1], as well as examining cross-cultural differences with respect to language use on social media [5]. However, these existing studies on contrastive opinion mining rely on an assumption that input data containing different opinion perspectives are separated into different collections beforehand. While this assumption might hold for some practical scenarios, quite often one needs to analyse contrastive opinion contained in a single collection such as text of streaming social media data.

In addition, it is natural that debates on some topics are more prominent or controversial than others, which indicates the importance of the topic. Therefore, being able to understand the prominence of a topic and the levels of contrastiveness of sentiment will enable one to quickly identify information that needs immediate attention. Finally, existing models generally interpret contrastive opinions solely in terms of the extracted topic words, which are not adequate to help us accurately understand the opinions presented in the corpus since the topic words only express shallow semantics. Therefore, it would be illuminating to consider the dependency between the sentences in the corpus and the topic of discussion in order to better understand and interpret contrastive opinion. The representative sentences also help to clarify the coherence of the extracted topics.

In this paper, we address the aforementioned issues by proposing a novel unified latent variable model (contraLDA) for mining contrastive opinion from text collections [6]. The proposed model makes several distinctive contributions, for it: (1) can be trained flexibly under weakly-supervised or fully-supervised settings, depending on the type of supervision information available; (2) automatically discovers contrastive opinion from both single and multiple text collections; (3) quantifies the strength of opinion contrastiveness towards the topic of interest, which could allow one to swiftly flag issues that require immediate attention; and (4) extracts sentences relevant to topics by adopting a strategy from [7], making sentiment-bearing topics clearer to users. Extensive experimental results show that our model outperforms several baseline models in terms of extracting coherent and distinctive sentiment-bearing topics which express contrastive opinions. The top sentences extracted by our approach further help us effectively understand and interpret sentiment-bearing topics. Lastly, we evaluate the performance our model in the supervised sentiment and topic classification task, in which contraLDA outperforms or gives comparable performance to five strong supervised baselines.

The rest of the paper is organised as follows. We first review the related work in §2, followed by detailed discussion of our model in §3. §4 and §5 present the experimental setup and results, respectively. Finally we conclude the paper in §6.

2 Related Work

2.1 Cross-collection opinion mining

There are several previous studies related to our work. Zhai et al. [8] introduced the Cross-Collection Mixture (ccMix) model, which is a probabilistic model for comparing text collections. The model extracts topics from comparable news sources and identified topics common to all the sources about a given event as well as topics that are unique to each news source. Similarly, the Cross-Collection LDA (ccLDA) model [4] extracted what is common to all the sources and what is unique to one specific source. The key difference between ccMix and ccLDA is that the former was built based on probabilistic latent semantic indexing (pLSI) while the latter is based on latent Dirichlet allocation (LDA). However, neither model considered modelling the opinions in text.

To bridge the gap, there are a number of works which address both topics and sentiments. The Multi-View Topic Model (mview-LDA) [9], which can be trained in both fully-supervised and semi-supervised settings, detects ideological bias at topic-level across multiple collections of data and presented summarised views from different opinion perspectives. Mukherjee and Liu [10] proposed several topic models for mining contentions from discussions and debates. Apart from discovering contention/agreement indicators, the proposed models also can model the interaction between authors and topics with regards to the reply-to relations and author-pair structures. Fang et al. [1] tackled the problem of mining contrastive opinions from political texts. They assumed that topics are expressed through nouns and opinions through adjectives, verbs, and adverbs. In addition, while opinion words are drawn from a perspective specific opinion distribution, topic words are drawn from a word distribution shared across multiple collections. The above assumption was also adopted by Thonet et al. [11], who proposed the Viewpoint and Opinion Discovery Unification Model (VODUM) model for joint discovery of viewpoints, topics and opinions. In their part-of-speech tagging process, nouns are excluded from the opinion word distribution, which potentially ignores words that are indicative for sentiment, e.g., $failure$, *genuis*, *wisdom*, etc.

Another line of work focuses on summarising contrastive opinion over multiple documents. Paul et al. [12] proposed a two-stage approach to summarising contrastive viewpoints. First, they extracted multiple viewpoints from text using the Topic Aspect Model (TAM) [13]; TAM was modelled based on the assumption that a word in a document belongs to either a topic, a viewpoint, both or neither. The second stage introduced the Comparative LexRank algorithm used for both ranking and generating contrastive summaries of the multiple viewpoints. Guo et al. [14] integrated expert opinions with ordinary opinions from social media for contrastive opinion summarisation, where the expert opinions were used as priors for aligning contrastive sentiment in the ordinary opinions. Ren and de Rijke [15] targeted contrastive theme summarisation using hierarchical non-parametric processes. They first employed a structured determinantal point process to extract a subset of diverse and salient themes, based on which the contrastive summaries were then generated using an iterative optimisation algorithm. A recent study [16] presented two differential topic models (dTM-Dirichlet and dTM-SAGE) for summarising the differences among document groups. The dTM-Dirichlet model captures unique word usage for each document group by modelling the group-specific word distribution, whereas, the dTM-SAGE model captures both group-specific topics and the unique characteristics of each document group as well as the background topics.

2.2 Cross-lingual and cultural analysis

Studies focus on cross-lingual and cross-cultural analysis are also closely related to our work. This form of analysis is useful for identifying the similarities and differences of opinion across different languages or cultures. Nakasaki et al. [17] proposed a topic model for visualising and analysing cross-lingual and cross-cultural differences from social blogs. They first created multilingual queries from Wikipedia entries for retrieving blog feeds. Next, statistical measures based on term probability and frequency were introduced for differentiating terms that are characteristic in one language or in both languages. Guo et al. [18] proposed the cross-lingual latent semantic association (CLaSA) model to learn and categorise words used to describe the same aspect of product features in different languages. In a similar vein, Elahi and Monachesi [5] used LDA to examine cross-cultural similarities and differences from social media data with respect to language use, but with a focus on analysing how two different cultures express emotions during romantic discussions on social media. Gutièrrez et al. [19] proposed a statistical model which learns common topics from multilingual and non-parallel data, and simultaneously discovers the different perspectives of the learned topics across the cultural groups.

To summarise, although the aforementioned models provide frameworks for mining contrastive opinion among different groups or sources, they all rely on the assumption that data containing different opinions are separated into different collections beforehand. However, this requirement might not be practical in real-world applications, for instance, detecting contrastive perspectives on certain topics where the input is streaming social media data. In addition, topics extracted by these models are opaque in terms of what sentences in the corpus express them, and thus could not help us gain deep insights of opinions encoded in the topics.

3 Methodology

We propose a model called contraLDA which offers a unified framework for mining contrastive opinions from text, where the source of text could be either a single collection or multiple collection of text. In addition, the contraLDA model can be trained flexibly under weakly-supervised or fully-supervised settings, depending on the type of supervision information available.

The graphical model of contraLDA is shown in Fig. 1. Given a collection of documents $\boldsymbol{D}$, assume that $\boldsymbol{D}$ can be divided in to C classes: $\boldsymbol{D} = \{D^c\}_{c=1}^{C}$ with D^c documents per class, each document d in class c is a sequence of N_d words, each word in the document is an item from a vocabulary with V distinct terms, and c is the class index. Also assuming that L and T are the total number of sentiment labels and topics, respectively, the complete procedure for generating a word w_n in contraLDA is as follows: first, one draws a topic z from

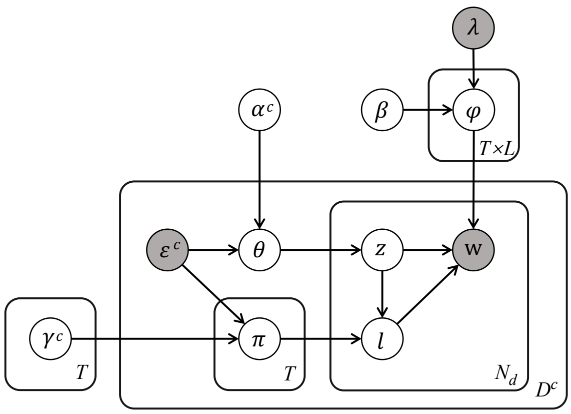


Fig. 1 The graphical model of contraLDA.

the class-constrained topic distribution $\boldsymbol{\theta}_d^c$. Following that, one draws a sentiment label l from the topic specific, class-constrained sentiment distribution $\boldsymbol{\pi}_{d,z}^c$. Finally, one draws a word from the per-corpus word distribution $\boldsymbol{\varphi}_{z,l}$ conditioned on both topic z and sentiment label l. Note that documents of all collections share the same $\boldsymbol{\varphi}$, and we can fully keep track of which collection a document belongs to based on its class index c. It is also important to note that the number of classes C plays a key role in controlling the operation mode of contraLDA. That is when $C = 1$, contraLDA is essentially modelling a single collection of text without any class membership information. In the scenario where $C > 1$, contraLDA will be switching to model multiple collections of text, e.g., documents annotated with class labels, or articles from New York Times and Xinhua News about the same set of events. We summarise the generative process of contraLDA as follows:

- For each topic $z \in \{1, \cdots, T\}$
  - For each sentiment label $l \in \{1, \cdots, S\}$
    - Draw $\boldsymbol{\varphi}_{z,l} \sim \mathrm{Dir}(\boldsymbol{\beta}_{z,l})$.
- For each document $d \in \boldsymbol{D}$
  - choose a distribution $\boldsymbol{\theta}_d^c \sim \mathrm{Dir}(\boldsymbol{\epsilon}_z^c \cdot \boldsymbol{\alpha})$.
  - For each sentiment label l under document d,
    - Choose a distribution $\boldsymbol{\pi}_{d,z}^c \sim \mathrm{Dir}(\boldsymbol{\epsilon}_l^c \cdot \boldsymbol{\gamma})$.
  - For each word $n \in \{1, \cdots, N_d^c\}$ in document d
    - Choose a topic $z_n \sim \mathrm{Mult}(\boldsymbol{\theta}_d^c)$,
    - Choose a sentiment label $l_n \sim \mathrm{Mult}(\boldsymbol{\pi}_{d,z_n}^c)$,
    - Choose a word $w_n \sim \mathrm{Mult}(\boldsymbol{\varphi}_{z_n,l_n})$.

3.1 Incorporating Supervised Information.

The contraLDA model can be trained flexibly under weakly-supervised or fully-supervised settings, depending on the type of supervision information available. Specifically, if there are only labelled features available (e.g., sentiment lexicon, or topic seed words), our model will incorporate the labelled features to constrain the Dirichlet prior of topic-word distributions, which essentially plays a role in governing the model inference. If there is fully labelled data available, e.g., labelled documents, our model will account for the full supervision from document labels during the generative process, where each document can associate with a single class label or multiple class labels. However, if the dataset contains both labelled and unlabelled data, our model will account for the available labels during the generative process as well as incorporate the labelled features as above to constrain the Dirichlet prior.

When labelled data is available, contraLDA incorporates supervised information by constraining that a training document can only be generated from the topic set with class labels corresponding to the document's observed label set. This is achieved by introducing a dependency link from the document label matrix $\boldsymbol{\epsilon}$ to the Dirichlet priors $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. Suppose a corpus has three topical labels denoted by $\boldsymbol{Z} = \{z_1, z_2, z_3\}$ and for each label z_k there are two sentiment labels denoted by $\boldsymbol{l} = \{l_1, l_2\}$. Given observed label matrix $\boldsymbol{\epsilon}_c = \{\boldsymbol{\epsilon}_z^c, \boldsymbol{\epsilon}_l^c\} = \{(1,0,1),(1,0)\}$ which indicates that d is associated with topic labels z_1, z_3 as well as sentiment label l_1, we can encode the label information into contraLDA as

$$\boldsymbol{\alpha}_d^c = \boldsymbol{\epsilon}_z^c \cdot \boldsymbol{\alpha} \tag{1}$$

$$\boldsymbol{\gamma}_d^c = \boldsymbol{\epsilon}_l^c \cdot \boldsymbol{\gamma} \tag{2}$$

This ensures that d can only be generated from topics associated with observed class labels from $\boldsymbol{\epsilon}$. If there are no labelled documents available, contraLDA will incorporate labelled features from $\boldsymbol{\lambda}$ (e.g., sentiment lexicons) for constraining the Dirichlet priors $\boldsymbol{\beta}$ using the same strategy described in [20, 21].

3.2 Inference

From the contraLDA graphical model depicted in Fig. 1, we can write the joint distribution of all observed and hidden variables which can be factored into three terms:

$$P(\mathbf{w}, \mathbf{z}, \mathbf{l} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{c}) = P(\mathbf{w}|\mathbf{z}, \mathbf{l}, \boldsymbol{\beta}) P(\mathbf{l}|\mathbf{z}, \boldsymbol{\gamma}, \boldsymbol{c}) P(\mathbf{z}|\boldsymbol{\alpha}, \boldsymbol{c}) \tag{3}$$

By integrating out $\boldsymbol{\phi}$, $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ in the first, second and third term of Eq. 3 respectively, we can obtain

$$P(\mathbf{w}|\mathbf{z}, \mathbf{l}, \boldsymbol{\beta}) = \int P(\mathbf{w}|\mathbf{z}, \mathbf{l}, \boldsymbol{\phi}) P(\boldsymbol{\phi}|\boldsymbol{\beta})\, d\boldsymbol{\phi} =$$

$$\prod_k \prod_j \frac{\Gamma(\sum_{i=1}^{V} \beta_{k,j,i}) \prod_i \Gamma(N_{k,j,i} + \beta_{k,j,i})}{\prod_{i=1}^{V} \Gamma(\beta_{k,j,i}) \Gamma(N_{k,j} + \sum_i \beta_{k,j,i})} \tag{4}$$

Algorithm 1 Sampling procedure for the contraLDA model.

Input: $\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}$, Corpus
Output: returns sentiment and topic label assignment for all word tokens, sentences and documents in the corpus
1: Initialize topic T, sentiment S, and word V matrices for $\boldsymbol{\theta}^c$, $\boldsymbol{\pi}^c$, and $\boldsymbol{\phi}$;
2: **for** $x = 1$ to max Gibbs sampling iterations **do**
3: **for** each documents $d \in \boldsymbol{D}$ **do**
4: **for** each word $w \in d$ **do**
5: Exclude word, sentiment label and topic at index x: $N^{\neg x}_{k,j,i}$, $N^{\neg x}_{k,j}$, $N^{\neg x}_{d,k,j}$, $N^{\neg x}_{d,k}$, $N^{\neg x}_{d}$
6: **if** $\boldsymbol{\epsilon}$ exists **then**
7: $\boldsymbol{\alpha}^c_d = \boldsymbol{\epsilon}^c_z \times \boldsymbol{\alpha}$
8: $\boldsymbol{\gamma}^c_d = \boldsymbol{\epsilon}^c_l \times \boldsymbol{\gamma}$
9: **else**
10: Sample a new sentiment-topic pair $\tilde{l}$ and $\tilde{z}$ using Equation 7;
11: **end if**
12: Update $N_{k,j,i}$, $N_{k,j}$, $N_{d,k,j}$, $N_{d,k}$ and N_d using $\tilde{l}$ and $\tilde{z}$ in step 5;
13: **end for**
14: **end for**
15: **for** every 25 iterations **do**
16: Update hyperparameter $\boldsymbol{\alpha}$ with maximum-likelihood estimation;
17: **end for**
18: **for** every 100 iterations **do**
19: Update matrices $\boldsymbol{\phi}$, $\boldsymbol{\theta}^c$, and $\boldsymbol{\pi}^c$ with new sampling results;
20: **end for**
21: **end for**

$$P(\mathbf{z}|\alpha, \mathbf{c}) = \int P(\mathbf{z}|\boldsymbol{\theta}, \mathbf{c})\, P(\boldsymbol{\theta}|\alpha, \mathbf{c})\, d\boldsymbol{\theta} = \prod_{c=1}^{C} \prod_{d=1}^{D^c} \frac{\Gamma(\sum_{k=1}^{T} \alpha^c_{d,k}) \prod_k \Gamma(N_{d,k} + \alpha^c_{d,k})}{\prod_{k=1}^{T} \Gamma(\alpha^c_{d,k}) \Gamma(N_d + \sum_k \alpha^c_{d,k})}, \tag{5}$$

$$P(\mathbf{l}|\mathbf{z}, \boldsymbol{\gamma}, \mathbf{c}) = \int P(\mathbf{l}|\mathbf{z}, \mathbf{c}, \boldsymbol{\pi})\, P(\boldsymbol{\pi}|\boldsymbol{\gamma}, \mathbf{c})\, d\boldsymbol{\pi} = \prod_{c=1}^{C} \prod_{d=1}^{D^c} \prod_{k} \frac{\Gamma(\sum_{j=1}^{L} \gamma^c_{d,k,j}) \prod_j \Gamma(N_{d,k,j} + \gamma^c_{d,k,j})}{\prod_{j=1}^{L} \Gamma(\gamma^c_{d,k,j}) \Gamma(N_{d,k} + \sum_j \gamma^c_{d,k,j})} \tag{6}$$

where $N_{k,j,i}$ is the number of times word i appeared in topic k with sentiment label j, $N_{k,j}$ is the number of times words are assigned to topic k and sentiment label j, $N_{d,k,j}$ is the number of times a word from document d is associated with topic k and sentiment label j, $N_{d,k}$ is the number of times topic k is assigned to some word tokens in document d, N_d is the total number of words in document d and Γ is the gamma function.

The main objective of inference in contraLDA is then to find a set of model parameters that can best explain the observed data, namely, the class-constrained topic proportion $\boldsymbol{\theta}^c$, the class-constrained topic label specific sentiment proportion $\boldsymbol{\pi}^c$, and the per-corpus word distribution $\boldsymbol{\varphi}$. To compute these target distributions, we need to calculate the posterior distribution of the model. As the posterior is intractable, we use a collapsed Gibbs sampler to approximate the posterior based on the full conditional distribution for each word token in position t. By evaluating the model joint distribution in Eq. 3, we can yield the full conditional distribution as follows

$$P(z_t = k, l_t = j|\mathbf{w}, \mathbf{z}^{-t}, \mathbf{l}^{-t}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{c}) \propto \frac{N^{-t}_{k,j,w_t} + \beta_{k,j,i}}{N^{-t}_{k,j} + \sum_i \beta_{k,j,i}} \cdot \frac{N^{-t}_{d,k} + \alpha^c_{d,k}}{N^{-t}_{d} + \sum_k \alpha^c_{d,k}} \cdot \frac{N^{-t}_{d,k,j} + \gamma^c_{d,k,j}}{N^{-t}_{d,k} + \sum_j \gamma^c_{d,k,j}}. \tag{7}$$

Using Eq. 7, we can obtain sampling assignments for contraLDA model, based on which model parameters can be estimated as

$$\varphi_{k,j,i} = \frac{N_{k,j,i} + \beta_{k,j,i}}{N_{k,j} + \sum_i \beta_{k,j,i}}, \tag{8}$$

$$\theta^c_{d,k,j} = \frac{N_{d,k} + \alpha^c_{k,j}}{N_d + \sum_k \alpha^c_{d,k}}, \tag{9}$$

$$\pi^c_{d,k} = \frac{N_{d,k,j} + \gamma^c_{d,k,j}}{N_{d,k} + \sum_j \gamma^c_{d,k,j}}. \tag{10}$$

3.3 Hyperparameter estimation

For the contraLDA model hyperparameters, while the values of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are set empirically, $\boldsymbol{\alpha}$ is estimated from data using maximum-likelihood.

Setting $\boldsymbol{\alpha}^c$. A common practice for topic model implementation is to use symmetric Dirichlet hyperparameters. However, it was reported that using an asymmetric Dirichlet prior over the per-document topic proportions has substantial advantages over a symmetric prior [22]. We initialise the asymmetric $\boldsymbol{\alpha} = (0.05 \times \bar{N})/T$, where $\bar{N}$ is the average document length of the corpus and the value of 0.05 on average allocates 5% of probability mass for mixing. Afterwards for every 40 Gibbs sampling iterations, $\boldsymbol{\alpha}$ is learned directly from data using maximum-likelihood estimation [22, 23]:

$$(\alpha^c_{z,l})^{\text{new}} \leftarrow \frac{\alpha^c_{z,l} \sum_d [\Psi(N_{d,z,l} + \alpha^c_{z,l}) - \Psi(\alpha^c_{z,l})]}{\sum_d [\Psi(N_{d,l} + \sum_{l'} \alpha^c_{z,l'}) - \Psi(\sum_{l'} \alpha^c_{z,l'})]}. \tag{11}$$

Setting $\boldsymbol{\beta}$. The Dirichlet prior $\boldsymbol{\beta}$ is first initialised with a symmetric value of 0.01 [22], and then modified by a transformation matrix $\boldsymbol{\lambda}$ which encodes the supervised information from the labelled feature learned from the training data.

Setting γ^c. We empirically set the symmetric prior $\gamma^c = (0.05 \times \bar{N})/(T \times L)$, where the value of 0.05 on average allocates 5% of probability mass for mixing.

3.4 Modelling the associations between sentiment-bearing topics and sentences.

Existing models can only learn topic-word and topic-document associations as they operate on bag-of-words features at the document-level, with the sentential structures of the corpus being ignored. Therefore, we adopt a computational mechanism [7] that can uncover the association between a sentiment-bearing topic and the underlying sentences of a corpus. First, we preserve the sentential structure of each document during the corpus preprocessing step (see §4 for more details). Second, modelling topic-sentence relevance is essentially equivalent to calculating the probability of a sentence given a sentiment-bearing topic $p(\text{sent}|z, l)$. The posterior inference of our model, based on Gibbs sampling, can recover the hidden sentiment label and topic label assignments for each word in the corpus. Such label-word assignment information provides a means for re-assembling the relevance between a word and a sentiment-bearing topic. By leveraging the sentential structure information and gathering the label assignment statistics for each word of a sentence, we can derive the probability of a sentence given a sentiment-bearing topic as

$$p(\text{sent}|z, l) = \frac{p(z, l|\text{sent}) \cdot p(\text{sent})}{p(z, l)} \propto p(z, l|\text{sent}) \cdot p(\text{sent}), \tag{12}$$

where

$$p(z, l|\text{sent}) = \frac{\sum_{w,z',l'} \varphi_{z',l',w}}{\sum_{w \in \text{sent}} \varphi_{z',l',w}}, \tag{13}$$

$$p(\text{sent}) = \sum_{z} \sum_{l} \prod_{w \in \text{sent}} \varphi_{z,l,w}. \tag{14}$$

Note that $p(l, z)$ is discounted as it is a constant when comparing sentential labels for the same sentiment-bearing topic. The extracted sentences for each sentiment-bearing topic are ranked based on their probability scores.

4 Experimental Setup

4.1 Dataset

We evaluate the performance of our model for contrastive opinion mining on two datasets with distinctive characteristics: (1) the Obama Healthcare dataset[1)] and (2) the El Capitan dataset[2)] [24].

Obama Healthcare dataset. This dataset contains telephone interview responses of 1,014 adults regarding the Obama Healthcare bill, out of which 45% of the responses are *for* the bill and 48% *against*[3)]. We choose this dataset because it has been widely used in many (contrastive) opinion mining related studies [2, 15].

El Capitan dataset. The El Capitan dataset consists of reviews manually annotated (with 18 topic labels and 3 sentiment labels in total) for various opinion mining tasks. The dataset consists of 2,232 customer reviews, with topic and sentiment annotations at both the review and sentence levels. For the sentiment labels, we only concentrate on positive and negative sentiment labels with the 2.3% of neutral reviews being ignored, since the aim of this study is to mine contrastive opinion from text.

4.2 Preprocessing

We preprocessed the experimental datasets by first performing automatic sentence segmentation[4)] in order to preserve the sentential structure information of each document. We then remove punctuation, numbers, non-alphabet characters, stop words, lowercase all words, and perform stemming. Summary statistics of the datasets are shown in Table 2.

4.3 Baselines

There are a few lines of study on contrastive opinion mining and viewpoint detection from textual data, which share the spirit of the proposed contraLDA model. We describe below the most relevant models which we employ as baselines in our experiment.

TAM model. The Topic Aspect Model (TAM) [13] jointly discovers topics and aspects which represent opinion perspectives. In the generative process of TAM, topic and aspect mixtures are sampled independently. In contrast, contraLDA models the dependency between topics and opinions and samples sentiment and topic labels simultaneously.

ccLDA model. The Cross-Collection LDA (ccLDA) model [4] detects the similarities and differences in topics between cultures from comparable blogs and forums. ccLDA assumes that the opinion perspective of a document is a known priori

1) http://www.gallup.com/poll/126521/favor-oppose-obama-healthcare-plan.aspx
2) https://github.com/eibeke/El-Capitan-Dataset
3) NB: the remaining 7% neutral responses are ignored.
4) http://www.nltk.org/

Table 2 Dataset statistics.

Dataset		Documents		Sentences		# of Words	Vocab. size
		Total num.	Avg. length	Total num.	Avg. length		
Obama Healthcare	For	434	16.4	574	12.6	6,577	1,188
	Against	508	17.1	684	12.8	7,719	1,395
El Capitan		2,232	80	10,348	17.3	178,668	17,873

since perspectives are determined by the collection a document belongs to.

VODUM model. The Viewpoint and Opinion Discovery Unification Model (VODUM) model [11] jointly discovers viewpoints, topics, and opinions from texts in an unsupervised setting. VODUM requires a bimodal dataset in which topical words and viewpoint specific opinion words are partitioned using part-of-speech (POS). Unlike VODUM, contraLDA and other baselines do not require POS tagging in preprocessing.

5 Experimental Results

In this section, we evaluate the performance of our model for contrastive opinion mining based on the two datasets described above. For the results reported for topic coherence (§5.1), contrastive opinion analysis (§5.2), and opinion contrastivenss (§5.3), contraLDA is trained with weakly-supervised learning as all the baseline models are weakly-supervised. For the classification results reported in §5.4, contraLDA is trained with fully-supervised learning.

5.1 Topic coherence

We first quantitatively measure the coherence of the extracted topics by our model and compare the results against a number of baselines, namely, TAM [13], ccLDA [4], and VODUM [11]. Topic coherence is a metric for measuring the quality of the extracted topics. This metric, in contrast to perplexity and likelihood, has been shown to be highly consistent with human experts in the task of assessing topic quality [25, 26]. Specifically, we employ normalised pointwise mutual information (NPMI) [27] to measure the semantic coherence of topics as it has been shown in a number of studies that NPMI outperforms other metrics for measuring topic coherence [28, 29]. Formally, NPMI is defined as

$$\mathrm{NPMI}(w_i, w_j) = \frac{\mathrm{PMI}(w_i, w_j)}{-\log(p(w_i, w_j))}, \tag{15}$$

where

$$\mathrm{PMI}(w_i, w_j) = \log_2 \frac{p(w_i, w_j)}{p(w_i)p(w_j)}. \tag{16}$$

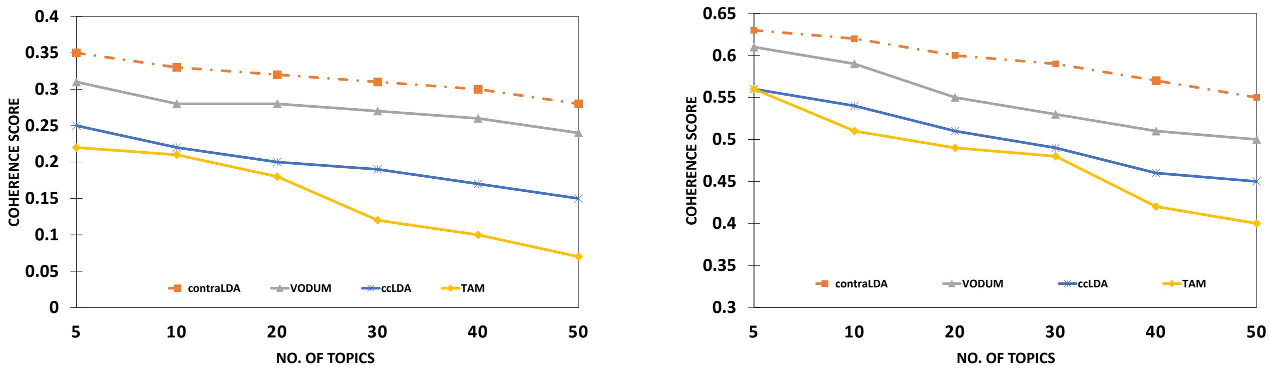
For both experimental datasets, we run our model and the baseline models with two sentiment labels (i.e., positive and negative), and vary the topic number setting $T \in \{5, 10, 20, 30, 40, 50\}$ (in a weakly supervised setting). For fair comparison, we set our model and all baseline models[5)] with the same topic-word distribution hyperparameter, i.e., $\beta = 0.01$ [30]. The document-topic distribution hyperparameter α is set to 0.1 for all models according to the original default setting described in [13]. For each model, we ran Gibbs sampling 5 times with 1000 iterations each run. We then average the topic coherence scores for each model over those 5 independent runs, as reported in Fig. 2.

As can be seen from Fig. 2, there is a general pattern for all tested models, where the coherence score of the extracted topics decreases as a larger number of topics K being modelled. This is inline with the observations of [19, 26], who discovered that as the number of topics increases, lower-likelihood topics tend to be more incoherent, resulting in lower coherence score for topics. It is also observed that topics extracted from the El Capitan dataset are more coherent than the topics from the Healthcare dataset. This is likely due to the fact that documents of the Heathcare dataset are much shorter than that of the El Capitan dataset (cf. Table 2), i.e., in short documents, word co-occurrence patterns are more difficult to discover and hence resulting in less coherent topics. In terms of individual models, our model consistently achieves a higher coherent score than all baseline models. For instance, when compared with the best baseline VODUM, our model gives over 8% and 15% averaged improvement on the El Capitan dataset and the Healthcare dataset, respectively. This demonstrates the capability of the proposed contraLDA in extracting coherent and meaningful topics.

5.2 Contrastive opinion analysis

In this section, we qualitatively evaluate our model in the task of discovering contrastive opinions. For the Healthcare dataset, all models were trained with 5 topic and 2 sentiment

5) For the TAM model, β is the prior for document-aspect distributions and ω the prior topic-word distribution (which is equivalent to β in the other models). Therefore, we set $\omega = 0.01$ and use the default value for $\beta = 1.0$ following [13].



(a) Obama Healthcare Dataset

(b) El Capitan Dataset

Fig. 2 Topic coherence analysis using NPMI.

labels following [2]. In terms of the El Capitan dataset, models were trained with 18 topics and 2 sentiment labels.

5.2.1 Mining contrastive opinions from text

The top panel of Table 3 shows 6 contrastive opinion topic pairs extracted by our model for both datasets, with each topic represented by the top 10 topic words. Note that a topic pair such as (`Topic0+`, `Topic0-`), expresses contrastive opinions towards the same topic `Topic0`, with '+' and '-' indicating the topic sentiment orientation. For instance, the two topics under `Topic0` show contrastive opinions about the proposed healthcare bill. Topic words such as *need*, *better*, *afford* extracted by contraLDA model indicate support for the bill, whereas words such as *debt*, *cost*, *control* show people's concern towards the bill. `Topic2` is likely about the effect of the bill on the country's economy, in which we see `Topic2+` seems to convey arguments from the *for* group that the bill is *good* and *better*, whereas by inspecting `Topic2-`, the *against* group is likely to hold reservation about this idea (i.e., *bad* and *expensive*). The right panel of Table 3 presents the contrastive opinion topics extracted from the El Capitan dataset, which contains a lot of highly controversial opinions regarding this Mac operating system[6]. For instance, the topic `Performance+` suggests that some people feel the system *performs better* and *app runs faster*, whereas the negative topic `Performance-` seems to show highly contrastive opinion that people have bad experience after upgrade, e.g., *app crashes* or *freezes*, and *mac* becomes *slow*.

For comparison, we also show 6 contrastive opinion topic pairs extracted by the VODUM model, i.e., the baseline model with the best topic coherence score (cf. Fig. 2). The topics extracted by VODUM, as shown in the bottom panel of Table 3, were aligned with the topics extracted by contraLDA automatically using pointwise mutual information [28,31]. It is observed that the contrastive topic pairs extracted by VODUM are either not coherent enough (i.e., difficult to interpret) or contain overlapping sentiment words, which make the viewpoints difficult to differentiate. For instance, both topics under `Performance` contain negative sentiment words like *slow* and *old*. In contrast, the sentiment-bearing topics extracted by our model are much more distinctive and crisp, which convey clear viewpoints.

5.2.2 Extracting relevant sentences for supporting opinion understanding

Although the topics discovered by contraLDA are meaningful and convey plausible contrastive opinions, it is still impossible to accurately interpret the meaning of the extracted topics solely based on its multinomial distribution, especially when one is unfamiliar with the topic domain. For example, topic words such as *crash*, *slow* and *freeze* under topic `Performance-` of the contraLDA model express clear negative sentiment. However, it is impossible to tell whether the sentiment is targeted to the aspect word *app* or *mac*. To address this gap and to gain deeper insight to the opinion encoded in the topics, we employed a mechanism [7] (as described in §3.4) to extract the most relevant sentences for sentiment-bearing topics, which can facilitate accurate understanding and interpretation of the topics discovered.

Table 4 shows the extracted top sentences (ranked based on Eq. 12) for each sentiment-bearing topic. For instance, the extracted top sentences for `Topic2` "Updating will be *good* for the *economy*" and "I think it's a detriment to the *economy*" show that topic words under `Topic2+` suggest that the Obama Healthcare will help the economy, while `Topic2-` contrastively indicates that the proposed health-

[6] NB: `Performance`, `Office` and `Yosemite` are label information from the El Capitan dataset.

Table 3 Contrastive opinion topic examples and the top rated sentence for each topic. Left: Healthcare dataset; Right: El Capitan.

Healthcare						El Capitan					
Topic0		Topic2		Topic3		Performance		Office		Yosemite	
+	-	+	-	+	-	+	-	+	-	+	-
contraLDA											
healthcar	healthcar	economi	compani	insur	cost	work	crash	offic	offic	yosemit	yosemit
peopl	govern	compani	economi	healthcar	uninsur	run	work	microsoft	use	work	upgrad
need	money	countri	dont	compani	insur	perform	time	compat	work	time	destroy
countri	control	good	healthcar	work	peopl	faster	app	quick	microsoft	downgrade	slow
better	involv	better	poor	peopl	chang	app	use	fine	ms	restor	work
everybodi	debt	need	worse	pay	increas	smooth	slow	work	crash	issu	mac
provid	owe	help	bad	health	health	new	mac	updat	issu	instal	bad
afford	moni	way	expens	believ	america	pro	open	upgrad	word	machin	problem
system	cost	go	dollar	need	debt	macbook	freez	new	excel	macbook	maverick
insur	person	think	high	think	expens	better	just	didn	appl	revert	appl
VODUM											
healthcar	afford	want	economi	insur	think	pro	pro	updat	work	new	new
think	need	think	american	peopl	go	updat	run	work	updat	like	like
go	think	economi	agre	realli	medicar	el	upgrad	use	use	great	realli
need	healthcar	compani	expens	high	just	instal	el	open	just	yosemit	updat
don	abl	pre	pay	got	read	run	updat	app	open	el	yosemit
better	like	exist	involv	like	understand	upgrad	new	upgrad	el	better	great
good	expens	public	go	won	insur	perform	instal	just	instal	updat	use
like	get	need	just	big	elderli	old	slow	like	run	make	good
help	realli	abl	think	help	believ	slow	work	el	upgrad	use	os
pay	don	like	save	hospit	want	late	old	slow	app	os	love

Table 4 Top sentences extracted based on the contraLDA model. NB: italic denotes words also appear in the corresponding topic.

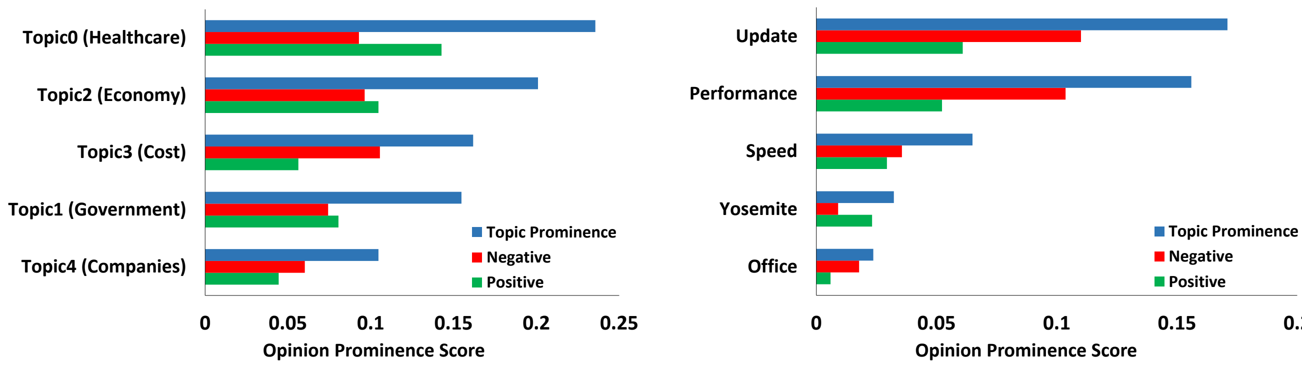
Healthcare		El Capitan	
Topic0 +	We *need affordable healthcare* for *everyone*.	Performance +	So much *better* than before, and *apps* run *faster* too.
Topic0 -	It's going to drive the *cost* of our *healthcare* up.	Performance -	Computer *slows* down dramatically, programs *freeze*.
Topic2 +	Updating will be *good* for the *economy*.	Office +	*Office* 2016 opens *quickly* with no issues.
Topic2 -	I think it's a detriment to the *economy*.	Office -	Update:*Office* apps tend to *crash* after the update!
Topic3 +	Need to provide *insurance* for the uninsured *people*.	Yosemite +	So I *downgraded* back to *Yosemite* and - hey presto!
Topic3 -	It doesn't address the *cost* of *insurance*.	Yosemite -	My 2010 *iMac* was *destroyed* by *Yosemite*.

care system will be detrimental to the economy. The top sentences for the `Office` topic show that some customers recorded an improvement with their office app (e.g., "*Office* 2016 opens *quickly* with no issues"), while others are unhappy with the office app (e.g., "Update: *Office* apps tend to *crash* after the update"). In a similar vein, the extracted sentences for the `Performance` topic help to clarify that while some people have an update with a system performance boost, others witnessed a dramatic performance drop. The extracted top sentences also help reveal the dependencies between topic words, which are essential for better interpretation of the topic. For instance, the sentence extracted for `Performance-` reveals that the sentiment word *freeze* actually describes the aspect word *apps*, and that the sentiment *slow* is directed to the computer, i.e., "Computer *slows* down dramatically, *apps freeze*".

To summarise, the top sentences extracted by our approach can effectively bridge the gap between the topic word distributions and the opinion encoded within the topic, and hence can greatly help facilitate sentiment-bearing topic understanding and interpretation.

5.3 Analysis of opinion contrastiveness

In the previous section, we qualitatively analyse the contrastive opinion topic pairs extracted by the contraLDA model. In this section, we further study the problem of quantifying the strength of opinion contrastiveness towards the topic of interest. We approach this by computing the prominence score for each sentiment-bearing topic extracted by contraLDA given a corpus $\boldsymbol{c}$. The prominence score is de-



(a) Obama Healthcare Dataset (b) El Capitan Dataset

Fig. 3 Analysis of topic prominence and sentiment contrastiveness. NB: blue bar indicates the overall prominence of contrastive topic pair; green bar indicates the strength of a positive sentiment topic, and red bar for negative sentiment topic.

fined as

$$P(z,l) = \frac{1}{|D|}\sum_{d=1}^{D} P(l|z,d)P(z|d) = \frac{1}{|D|}\sum_{d=1}^{D} \theta_{d,z} \cdot \pi_{d,z,l}, \quad (17)$$

where D is the total number of documents in the corpus. Thus the prominence for topic z in a corpus can be derived as

$$P(z) = \sum_{l} P(z,l). \quad (18)$$

Fig. 3 shows some contrastive opinion topic pairs ordered by their prominence in the corpus, where the pairs with the highest prominence scores (calculated based on Eq. 18) are placed on the top. One of the benefits of modelling topic prominence and sentiment contrastiveness is that it gives a quick overview of the notable topics and the sentiments towards them, which allows one to swiftly flag topics or issues that require immediate attention. For instance by looking at Fig. 3, one can easily identify that for the Healthcare dataset, `Topic0` and `Topic2` gained most concerns from the public, whereas for the El Capitan dataset the most heated topics are `update` and `performance`.

In terms of opinion contrastiveness, we see that `Topic2`, which conveys opinion about the effect of the proposed healthcare system on the economy, received quite balanced positive and negative sentiment magnitude. In contrast, for `Topic3`, negative opinions on the cost of insurance significantly outweigh positive opinions. By inspecting the topic words and the representative sentences in the corpus, many people expressed concern about the cost, e.g., "It's going to cost us too much and won't cover what is expected." In terms of the El Capitan dataset, we can see that topics `Performance` and `Update` are skewed towards the negative sentiment, indicating that a majority of customers experienced a performance drop after upgrading to El Capitan. Interestingly, for the `Yosemite` topic, positive sentiments toward the topic clearly outweighs the negative sentiments. By examining the corpus, we found that many people actually expressed that they preferred using Yosemite compared to the unstable El Capitan, e.g. "So I downgraded back to Yosemite".

5.4 Sentiment and topic classification

Our last experiment focuses on the task of sentiment and topic classification. When performing classification, training and testing data are normally prepared in the same modality, i.e., both at the document or the sentence level. In this experiment, we further explore the effect on classification performance when the data modality for training and testing are different based on the El Capitan dataset, i.e., train on documents and test on sentences, and vice versa. Note that sentiment classification in our experiment is a binary classification task involving positive and negative labels only. Topic classification is multi-class classification with 18 different labels. Since the baselines used in the previous section are weakly-supervised, we compare the overall performance of contraLDA (in supervised setting) with two supervised topic models (i.e. labelled LDA [32] and supervised LDA [33]), the Naive Bayes model (NB), SVM, as well as a deep learning baseline i.e. Convolutional Neural Network for Sentence Classification (CNN-SC) [34]. Recall, Precision and F1 score are used as evaluation metrics, and we report the results based on 5-fold cross validation, which are summarised in Tables 5 and 6.

Overall, it can be seen from Tables 5 and 6 that both NB and SVM consistently outperform the supervised topic

Table 5 Sentiment classification results on the El Capitan dataset. (NB: data in bold indicates the best overall results.)

Trained on	Tested on	Naïve Bayes			SVM			sLDA			L-LDA			CNN-SC			contraLDA		
		Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Sentence	Sentence	79.4	77.8	78.6	76.2	75.1	75.6	74.0	75.0	74.5	72.5	67.5	69.9	79.0	78.4	78.7	80.0	80.2	**80.1**
	Review	89.5	89.2	89.3	88.1	87.5	87.8	79.2	78.8	79.0	74.5	70.3	72.3	89.3	88.9	89.1	90.4	89.9	**90.1**
Review	Sentence	81.2	80.6	**80.9**	74.4	72.8	73.6	70.9	72.3	71.6	77.8	75.6	76.7	80.1	79.8	80.0	76.3	75.9	76.1
	Review	84.8	82.8	83.8	75.9	73.9	74.9	78.5	78.6	78.5	84.0	82.6	83.3	84.6	85.4	**85.0**	84.9	84.4	84.6

Table 6 Topic classification results on the El Capitan dataset. (NB: data in bold indicates the best overall result.)

Trained on	Tested on	Naïve Bayes			SVM			sLDA			L-LDA			CNN-SC			contraLDA		
		Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Sentence	Sentence	49.7	50.1	49.9	60.6	57.3	58.9	34.1	37.9	35.9	43.5	42.3	42.9	72.0	70.0	**71.0**	48.1	48.3	48.2
	Review	58.3	53.2	55.6	47.7	39.3	43.1	38.0	32.6	35.3	49.6	48.4	49.0	58.5	58.1	58.3	64.6	54.3	**59.0**
Review	Sentence	44.3	35.8	33.0	58.5	37.4	37.3	24.3	22.9	23.6	40.0	37.7	38.8	52.0	51.6	**51.8**	42.5	42.0	42.2
	Review	32.9	33.7	33.3	45.8	46.8	45.3	27.9	33.5	30.4	36.1	34.8	35.4	56.2	57.1	**56.7**	37.0	37.9	37.4

model baselines (i.e, sLDA and L-LDA) for both sentiment and topic classification. In terms of NB and SVM, it is observed that NB performed better than SVM in sentiment classification for all settings, with a higher margin from 2.7% to 8.8% in F1. However, SVM generally outperformed NB on topic classification, especially when both trained and tested at the sentence level. The proposed contraLDA outperformed all baselines in sentiment classification when trained on sentences. For topic classification, contraLDA achieved the best classification results when trained on sentences and tested on reviews. The deep learning approach (CNN-SC) achieved the best overall performance for topic classification. Especially, when trained on sentences and tested on sentences, CNN-SC gives a much higher F1 score (i.e., 71.0%) than all other comparison models, demonstrating its superior capability in handling large numbers of classes in classification.

One interesting finding from the experiment is that training and testing on the same data modality does not necessary yield the best classification result. In fact, for both sentiment and topic classification, most of the tested classifiers in general performed best when trained on sentences. For instance, contraLDA trained on sentences achieved the highest F1 scores in sentiment classification, i.e, 80.1% when tested on sentences and 90.1% when tested on reviews. In terms of topic classification, models trained on sentences again achieved the best results, i.e., contraLDA performed best when predicting review labels (i.e., 59.0%) and CNN-SC performed best when predicting sentence labels (i.e., 71.0%). These observations suggest that training data at the sentence level has better feature representation for class labels, as training data at the review level is more likely to introduce noisy features. For instance, even though the overall sentiment of a review is positive, it is not unusual that it also contain sentences expressing negative sentiment. Likewise, while a sentence normally discuss a certain topic, a review is likely to cover several topics or aspects [35].

Overall, it can be concluded that the proposed contraLDA model gives competitive performance in both sentiment and topic classification tasks, and models trained on sentences yield better performance than trained on reviews in general.

6 Conclusion

In this paper, we presented the contraLDA model for automatically mining contrastive opinion from either single or multiple text collections. Experimental results on two real-world datasets show that our model outperforms several commonly benchmarked models in extracting more coherent sentiment-bearing topic pairs which represent contrastive opinions. In addition, we introduced a mechanism for extracting sentences from corpus that are relevant to sentiment-bearing topics, which helps understanding and interpretation of the topics discovered. Apart from qualitatively analysis, we also quantitatively analysed the level of opinion contrastiveness towards topics of interest, which could allow one to swiftly flag issues that require immediate attention. Lastly, we evaluate the performance our model in the supervised sentiment and topic classification task, in which contraLDA outperforms or gives comparable performance to five strong supervised baselines.

In the future, we plan to investigate our approach on datasets from more domains. Another interesting direction is to extend the model with online learning capability, so that the model can fit large scale of data efficiently.

Acknowledgements This work is supported by the award made by the UK Engineering and Physical Sciences Research Council (Grant number: EP/P005810/1).

References

1. Y. Fang, L. Si, N. Somasundaram, and Z. Yu. Mining contrastive opinions on political texts using cross-perspective topic model. In *Proceedings of the International Conference on Web Search and Data Mining (WSDM)*, pages 63–72, 2012.
2. A. Trabelsi and O. R. Zaïane. A joint topic viewpoint model for contention analysis. In *Natural Language Processing and Information Systems (NLPIS)*, pages 114–125. 2014.
3. K. Lerman and R. McDonald. Contrastive summarization: an experiment with consumer reviews. In *Proceedings of human language technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 113–116, 2009.
4. M. Paul and R. Girju. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 1408–1417, 2009.
5. M. F. Elahi and P. Monachesi. An examination of cross-cultural similarities and differences from social media data with respect to language use. In *The EAS Workshop in the 10th Language Resources and Evaluation Conference (LREC)*, pages 4080–4086, 2012.
6. E. Ibeke, C. Lin, A. Wyner, and M. H Barawi. Extracting and understanding contrastive opinion through topic relevant sentences. In *International Joint Conference on Natural Language Processing (IJCNLP), Taiwan*, 2017.
7. M. H. Barawi, C. Lin, and A. Siddharthan. Automatically labelling sentiment-bearing topics with descriptive sentence labels. In *In Proceedings of the 22nd International Conference on Natural Language and Information Systems (NLDB)*, pages 299–312, 2017.
8. C. Zhai, A. Velivelli, and B. Yu. A cross-collection mixture model for comparative text mining. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pages 743–748, 2004.
9. A. Ahmed and E. P. Xing. Staying informed: supervised and semi-supervised multi-view topical analysis of ideological perspective. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1140–1150, 2010.
10. A. Mukherjee and B. Liu. Mining contentions from discussions and debates. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pages 841–849, 2012.
11. T. Thonet, G. Cabanac, M. Boughanem, and K. Pinel-Sauvagnat. Vodum: a topic model unifying viewpoint, topic and opinion discovery. In *European Conference on Information Retrieval (ECIR)*, pages 533–545, 2016.
12. M. J. Paul, C. Zhai, and R. Girju. Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 66–76, 2010.
13. M. Paul and R. Girju. A two-dimensional topic-aspect model for discovering multi-faceted topics. *In Association for the Advancement of Artificial Intelligence (AAAI)*, 2010.
14. J. Guo, Y. Lu, T. Mori, and C. Blake. Expert-guided contrastive opinion summarization for controversial issues. In *Proceedings of the 24th International Conference on World Wide Web Companion (WWW)*, pages 1105–1110, 2015.
15. Z. Ren and M. de Rijke. Summarizing contrastive themes via hierarchical non-parametric processes. In *Proceedings of the 38th SIGIR International Conference on Research and Development in Information Retrieval*, pages 93–102, 2015.
16. L. He, W. Li, and H. Zhuge. Exploring differential topic models for comparative summarization of scientific papers. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1028–1038, 2016.
17. H. Nakasaki, M. Kawaba, T. Utsuro, and T. Fukuhara. Mining cross-lingual/cross-cultural differences in concerns and opinions in blogs. In *Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy*, pages 213–224. Springer, 2009.
18. H. Guo, Z. Zhu, H.and Guo, X. Zhang, and Z. Su. Opinionit: a text mining system for cross-lingual opinion analysis. In *Proceedings of the 19th ACM International Conference on Information and knowledge management*, pages 1199–1208, 2010.
19. E.D. Gutiérrez, E. Shutova, P. Lichtenstein, G. de Melo, and L. Gilardi. Detecting cross-cultural differences using a multilingual topic model. *Transactions of the Association for Computational Linguistics (TACL)*, 4:47–60, 2016.
20. C. Lin and Y. He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 375–384. ACM, 2009.
21. C. Lin, Y. He, R. Everson, and S. Ruger. Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data Engineering*, 24(6):1134–1145, 2012.
22. H. Wallach, D. Mimno, and A. McCallum. Rethinking lda: Why priors matter. volume 22, pages 1973–1981, 2009.
23. T. Minka. Estimating a Dirichlet distribution. Technical report, MIT, 2003.
24. E. Ibeke, C. Lin, C. Coe, A. Wyner, D. Liu, M. H. Barawi, and N. F. Abd. Yusof. A curated corpus for sentiment-topic analysis. In *The Emotion and Sentiment Analysis Workshop in the 10th Language Resources and Evaluation Conference (LREC), Slovenia*, 2016.
25. J. Chang, C. Gerrish, S.and Wang, J. L. Boyd-Graber, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 288–296, 2009.
26. D. Mimno, H. M. Wallach, E. Talley, and A. Leenders, M.and McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 262–272, 2011.
27. G. Bouma. Normalized (pointwise) mutual information in collocation extraction. *In The German Society for Computational Linguistics and Language Technology (GSCL)*, pages 31–40, 2009.
28. D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The An-*

nual Conference of the North American Chapter of the Association for Computational Linguistics (ACL), pages 100–108, 2010.
29. N. Aletras and M. Stevenson. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS)*, pages 13–22, 2013.
30. M. Steyvers and T. Griffiths. Probabilistic topic models. pages 424–440, 2007.
31. A. Cano Basave, Y. He, and R. Xu. Automatic labelling of topic models learned from twitter by summarisation. In *Proceedings of 52nd Annual Meeting of Association for Computational Linguists*, pages 618–624, 2014.
32. D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 248–256, 2009.
33. J. D. Mcauliffe and D. M. Blei. Supervised topic models. In *Advances in neural information processing systems*, pages 121–128, 2008.
34. Y. Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.
35. G. Ganu, A. Marian, and N. Elhadad. Ursa-user review structure analysis: Understanding online reviewing trends. 2010.

Ebuka Ibeke is currently a PhD student at the University of Aberdeen and a Part-time Lecturer at the School of Computing Science and Digital Media, Robert Gordon University. He received his BSc degree (2.1 division) in Computer Science from Nnamdi Azikiwe University, Nigeria in 2005 and an MSc degree (1st class Honours) in Information Engineering from Robert Gordon University, UK in 2011. His research interests include sentiment analysis, topic modelling, argumentation mining, and natural language generation/processing.

Chenghua Lin received the BEng degree in electrical engineering and automation from Beihang University (2006), the MEng degree (First class Honours) in electronic engineering from the University of Reading (2007), and the PhD degree in computer science from the University of Exeter (2011). Currently, he is a SICSA Senior Lecturer (Associate Professor) in Computing Science at the University of Aberdeen. His current research interests include integration of machine learning and natural language processing for sentiment analysis, intention mining, text summarisation, and natural language generation.

Dr. Wyner is a Lecturer in Computing Science at the University of Aberdeen. He has a PhD in Linguistics (Cornell University, 1994) on the formal syntax and semantics of adverbial modification as well as a second PhD in Computer Science (King's College London, 2008) on the representation and automation of legal concepts for e-contracting. He has worked as a research associate on two EU projects to formalise the law and argumentation to automatically process and to support policymaking. He is currently a co-investigator in a project to make historical legal texts machine readable. He has numerous publications on legal informatics, text analysis, language processing, and argumentation.

Mohamad Hardyman Barawi is currently a PhD student at the University of Aberdeen and a Lecturer at the University Malaysia Sarawak. Prior to the academic post, he was operation support specialist for Hewlett Packard Malaysia. He received his BSc degree (Hons) in Computer Science from University Putra Malaysia in 2003 and an MLIS degree in Library and Information Science from University of Malaya, Malaysia in 2006. His research interests include topic modelling, sentiment analysis, and text summarisation.