Topic-Sentiment Mining from Multiple Text Collections

Qifeng Zhu and Fang Li

Dept. of Computer Science and Engineering
Shanghai Jiao Tong University
Shanghai, China 200240

Abstract. Topic-sentiment mining is a challenging task for many applications. This paper presents a topic-sentiment joint model in order to mine topics and their sentimental polarities from multiple text collections. Text collections are represented with a mixture of components and modeled via the hierarchical Dirichlet process which can determine the number of components automatically. Each component consists of topic words and its sentiments. The model can mine topics with different proportions and different sentimental polarities as well as one positive and one negative topic for each collection. Experiments on two text collections from Chinese news media and microblog show that our model can find meaningful topics and their different sentimental polarities. Experiments on Multi-Domain Sentiment Dataset show that our model is better than the JST-alike models on parameter settings for topic-sentiment mining.

Keywords: text mining, topic modeling, sentiment analysis, hierarchical Dirichlet process

1 Introduction

With the rapid development of Web 2.0, detecting topics and sentiments from the explosion of digitalized text stream of different media or different collections has become an important task. When some event happens, traditional media and social media spread the information about the event very fast. People wish to know what are the topics and their sentimental polarities, not only from the traditional media but also from many social media. There are also many product reviews collections. People want to know what are the topics and their polarities, which are valuable information when they buy some products.

A class of probabilistic topic models [1] has been developed to automatically cluster, index and find those underlying topics from large text collections. A document is assumed as a mixture of components or topics in these mixture models, whereas a topic is represented as a distribution over words. Sentiment analysis [11] aims to reveal what people think about an event or a product and so on. It focuses not only on a sentence, or a document, but also on a collection.

This paper presents a model to represent topic-sentiment components, which uses the hierarchical Dirichlet process (HDP) [15] to mine topics and their corresponding sentiment polarities from multiple collections. For example, suppose that there are two product review collections; one is about DVDs and the other is about electronic products. Using the proposed model, one can answer:

- What are the main topics in these two collections?
- What are the sentiment polarities towards these topics?
- What are the differences among DVDs and Electronics reviews?

The proposed model has the ability to:

- Determine automatically the number of underlying topics for different text collections. Other topic models such as the latent Dirichlet allocation (LDA) [2] need to specify the number of topics before using it;
- Reduce the dimensions and pre-allocations of model parameters and latent variables. It does not need to allocate two or more dimensions as found in other previous works like [7, 9], where the number of topic will be not reasonable for some cases.
- Inference the topics and their sentiment polarities together. People can compare the differences of the topics and their sentimental polarities among multiple collections.

In the following, we first discuss about the related work, then our model and finally the experiments.

2 Related Work

Many researches focuses on topics and sentiments jointly, either mining or modeling [4, 7, 9, 10, 12, 13]. Among them, three kinds of models need to be mentioned here.

Different Viewpoints Discovery. For any issue, there are positive and negative viewpoints. Paul et al. [13] used a multi-faceted joint model to mine “for” and “against” viewpoints from opinionated text collections, such as editorials about the Israel-Palestine conflict. A document is assumed as a mixture of multi-dimensional components: a sentiment dimension and a topic (or aspect) dimensions. Fang et al. [5] proposed a model to regard topic and opinion as two aspects of multiple collections. Topics are the same across different text collections, while the opinions are specific to each collection. They set the same number of topics and opinions on different collections, while it is not always the same in reality.

Joint Sentiment Topic Model. Lin and He [9] proposed a joint sentiment topic model (JST) for unsupervised sentiment classification. It treats each component in the document as sentiment-topic-specific. Each sentiment has the same number of topics. A model selection problem was raised whether the sentiment-topic distribution shall be modeled as per-document distribution or per-collection distribution in that paper. The subsequent works [4, 7, 10] have made different choices. In this paper, the document-level topic distributions are generated by the collection-level ones using the HDP and thus it automatically makes the suitable decision for each document.

Topic Sentiment Mixture Model. Mei et al. [12] built a supervised model for mining sentiments associated with Weblogs. In the model a document is composed of several themes. Each theme consists of the positive, negative and neutral contents. A background component was also used to capture common words alongside these themes. They used a training collection with topic and sentiment labels for each document to learn the positive and negative sentiment models. Our model adopts the topic sentiment mixture with HDP, only uses sentiment lexicons as the sentiment prior, and treats the common words as a part of a component.

3 Our Model

A document is modeled as a mixture of several components. The number of components is determined by the underlying document collections. Instead of modeling a component as a Dirichlet-Multinomial for words like other topic models, our model treats the component itself as a mixture to be estimated. Each component consists of topic words, sentiment words, and some other (background) words in order to compose a human-readable document (named white noise). In our model each component consists of four parts: **topics**, **positive words**, **negative words**, and **white noise**. Each word is associated with a **component**, and a **label** to indicate a part of the component. A word in a document may serve as a topical word, a positive word, a negative word, or a white noise word. Sentiment words are recognized by a sentiment dictionary. In this case, the model can capture topics and sentiment polarity associated with topics.

We assume that different text collections refer to the same event, the same type of collections or things[1]. There are some assumptions for the model:

- Different text collections share the same topic words with different proportions (including zero) and different sentimental polarities towards each topic.
- Different collections have their own positive and negative word distributions.

3.1 Topic-Sentiment Model for Text Collections

Table 1 lists some important notations in this paper. Other notations for a model, such as the prior hyper-parameters, are omitted here for simplicity. They will be stated when they are used.

The model for two collections J_1 and J_2 is illustrated in Fig.1 with parameters. The generative process can be described as follows,

1. (a) draw a countable sequence of topics – $\{\theta_t\}$
   (b) draw positive word distributions, each collection owns one – $\{\theta_p\}$
   (c) draw negative word distributions, each collection owns one – $\{\theta_n\}$
   (d) draw a white noise word distribution – θ_o

[1] These collections may vary a lot in the content, but refer to similar topics.

Table 1. Major notations used in this paper

Notation	Meaning	Example
Topic θ_t	topic	love story dvd
Pos θ_p	positive words	cool great wonderful
Neg θ_n	negative words	bad terrible ugly
O θ_o	white noise	
δ	component	
c	collection index	$1, 2, \ldots$
j	document index	$1, 2, \ldots$
i	word index	$1, 2, \ldots$
π	component proportion	π_0, π_c, π_{cj}
τ	label proportion	$\tau_t, \tau_p, \tau_n, \tau_o$
z	component assignment	
l	label assignment	t, p, n, o

2. draw the proportions $\{\tau\}$ for each collection and form the components $\{\delta\}$ as a mixture using topics, positive words, negative words, white noise with the proportions $\{\tau\}$
3. draw a Dirichlet process over the components as the top level mixture distributions, emitting the proportion π_0 and the component number K
4. draw a Dirichlet process over π_0 for each collection c, emitting $\{\pi_c\}$
5. draw a Dirichlet process over π_c for each document in every collection, emitting $\{\pi_{cj}\}$
6. for each word in every document,
   (a) draw the component number and the label according to π_{cj} and τ jointly: z, l
   (b) draw the word according to the selected word distribution indexed by z, l.

Word distribution θs are modeled as Dirichlet distributions over word vocabulary with hyperparameters $\{\beta\}$. Positive and negative words specified by a sentiment lexicon $\{\lambda\}$. The positive words will not be generated by the negative sentimental topic and vice versa. Dirichlet prior is also applied to τ with hyperparmeters ς. The hierarchical Dirichlet process is parameterized by $\gamma, \alpha_0, \alpha_1$ respecting top, collection level and document level. The inference of this model consists of the posterior representation sampler of HDP [14, 15] and the resampling for z, l by their marginal distributions.

4 Experiments

4.1 Experimental Setup

Data Description. There are two kinds of datasets in our experiments. One dataset is the Multi-Domain Sentiment Dataset (MDS)[2] used for sentiment classification [3]. Each domain in the dataset is regarded as a document collection.

[2] `http://www.cs.jhu.edu/~mdredze/datasets/sentiment/` .

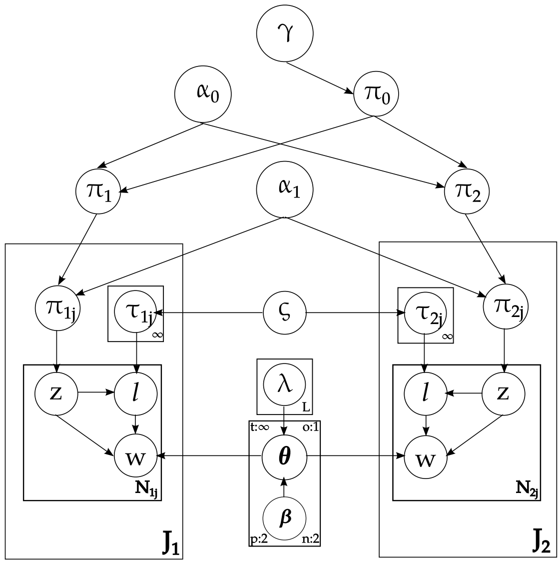

Fig. 1. Model for mining topics and their sentiment polarities (two collections)

DVD vs. Electronics domains and Games vs. Software domains are regarded as two multiple collections in the experiments. The detailed description is presented in Table 2. The other dataset is about two events from News and Sina Weibo in China. News collection was crawled from three Chinese news websites (Sina, iFeng, Tencent) during the end of 2011. Weibo collection was collected from Sina Weibo API using event keywords at the same time. There are two events; the first one ("Occupy Wall Street") happened in USA in the year of 2011, whereas the second one ("Little Yue Yue") happened in the south of China in the year of 2011. The detailed information of the dataset is shown in Table 3. Both events attracted many Chinese microblog users as well as traditional news media and lasted about two weeks. This dataset will be referred as EVENT. We use it for result presentation.

Table 2. The MDS domains used in experiments

	DVD			Electronics		
DVD vs Electronics	#	#Pos.	#Neg.	#	#Pos.	#Neg.
	1000	519	481	1000	488	512
	Games			Software		
Game vs Software	#	#Pos.	#Neg.	#	#Pos.	#Neg.
	561	483	78	585	300	285

Table 3. Event-related microblogs and news

Events	Microblogs		News Reports	
	#Blogs	#Words	#Reports	#Words
Occupy Wall Street	83K	2,250K	269	90K
Little Yue Yue	150K	4,200K	123	71K

Aim of the Experiments:

- whether the topics are correct for the text collections;
- whether the sentiment analysis works on the text collections.

Model Settings. The Dirichlet hyperparameter over words for each word distribution β is set to 0.01 as in [9, 10]. The Dirichlet hyperparameter ς is set as follows: the proportions among topic, sentiment, white noise have a uniform prior; the proportions between positive and negative have a Dirichlet(**0.5**). This makes the model capture the positive and negative words/sentiment of a document. The HDP hyperparameters $\gamma, \alpha_0, \alpha_1$ are chosen in range $[1, 10]$ empirically. We use the emotion ontology constructed by DUTIR[3] [16] for Chinese (regarding "happy", "like" as positive sentiments and "angry", "dislike", "fear", "sad" as negative sentiments) and the HowNet's words for sentiment analysis[4] for English as the sentiment lexicons $\{\lambda\}$.

4.2 Evaluation Metrics

In order to evaluate the correctness of topics mined from multiple collections, we design a human evaluation method. A topic is represented with the top 10 words. Each student was asked to give a score for each topic mined from the multiple collections: 2 for clear, valuable topics, 1 for somewhat meaningful topics, 0 for no meaning, but appear in these texts, -1 for incomprehensible ones and -2 for wrong ones. As stated in [6], negative values up to 0 suggest being a failure while positive scales suggest degree of success. They were asked to give a meaningful phrase for each topic as the golden standard. Seven students joined the evaluation process. The Krippendorff's α-value[8] among the seven evaluators is 0.68. We use two metrics: Grade and Accuracy. Grade is the average score for each topic evaluated by students. The accuracy is calculated in the following:

$$\text{Accuracy} = \frac{\#[g_k \geq 1]}{\#k} \quad (1)$$

, where k indices topics and g_k is the grade for each topic.

[3] `http://ir.dlut.edu.cn/EmotionOntologyDownload.aspx` (in Chinese).
[4] `http://keenage.com/html/c_bulletin_2007.htm` (in Chinese).

4.3 Topic Evaluation

The evaluation results are shown in Table 4. The average grade is above zero and hence it means some degree of success. “Game vs. Software” and “Little Yue Yue” have a grade 1, which means meaningful topics. The low accuracy for “DVD vs. Electronics” is due to the fact that there are topics like “get, make, give, look, feel” and “one, time, two, little, never”, which are considered incorrect by the human judges. The uniform prior among topic, sentiment, and white noise may be not suitable for that collection. For the event “Occupy Wall Street”, since there are topics such as advertisements of losing weight (Table 5 No.5) , the grade is low.

Table 4. Topic evaluation

Multi-Collection	#Topics	Grade	Accuracy
DVD vs. Electronics	6	0.50	0.50
Game vs. Software	9	1.0	0.67
Occupy Wall Street	9	0.24	0.56
Little Yue Yue	9	1.4	0.89

4.4 Topic-Sentiment Evaluation

We use the EVENT dataset for topic-sentiment evaluation. As the golden standard is not available, we represent topics and collection-specific sentiment words using top-10 words. The results are listed in Tables 5, 6 with English translation via Google. Each row represents a topic, the right two columns are the proportion and the sentiment for each topic, where ‘+’ for positive topics, ‘-’ for negative topics and ‘X’ for neutral ones for each collection. A topic k is considered to be positive (negative) in a collection c if the sum of the positive and negative proportion, $\tau_{c(p+n)}(k)$ is not zero and the ratio of the positive (negative) proportion to $\tau_{c(p+n)}(k)$ is greater than 0.6 (considering the estimation variance). Otherwise, the topic is considered as neutral.

Table 5 shows the result from the event of “Occupy Wall Street”. News reports focused more on the factual topics, such as No.4 and No.6 with the proportion of 0.241 and 0.234 respectively, both with the neutral sentimental polarity. On the other collection, Microblogs had the positive polarity towards topics No. 3, No. 4, No. 6 and No. 7. Topics of No. 1, 5, 7 and 9 are not in the News reports since their topic proportions are 0. The last four rows are positive and negative topics mined from news and microblogs. The positive topic of Microblog is “revolution”, “support” and so on. The positive topic of news reports is “support”, “development”, “response” and so on.

Table 6 shows the result from the event of “Little Yue Yue”. Microblog focused on the event topic with No. 2(“incident”), No. 3(“candles”) and No.7(“care”) because this event was first disseminated through the Microblogging platform.

Table 5. Topics mined from "Occupy Wall Street"

No.	Topics Mined in "Occupy Wall Street"	Microblog	News
1	诺基亚, 倾世, 华尔街, 惊心, 吴奇隆, 旅行, 微博去, 加油站, 皇妃, 肯德基 (Nokia, dumping the World, Wall Street, startling, Nicky, travel, go microblogging, gas stations, Princess, Kentucky)	0.079/X	0/X
2	运动, 金融, 社会, 政府, 认为, 民众, 总统, 人们, 政治, 冻结 (sports, financial, social, government, believe, the people, the president, people, politics, freezing)	0.055/X	0.085/X
3	美国, 华尔街, 占领, 国家, 金融, 问题, 人民, 没有, 世界, 政府 (United States, Wall Street, occupation, country, financial, issues, people, no, world, government)	0.188/+	0.196/X
4	华尔街, 占领, 中国, 美国, 知道, 公司, 报道, 人民, 运动, 已经 (Wall Street, occupation, China, the United States, knowing, company, reported, the people, the movement, already)	0.297/+	0.241/X
5	华尔街, 减肥, 乔布斯, 分享, 综合症, 命运, 郑州, 诺贝尔, 传说, 非常 (Wall Street, lose weight, Jobs, sharing, syndrome, fate, Zhengzhou, Nobel, legends, very)	0.088/X	0/X
6	活动, 纽约, 组织, 城市, 记者, 民众, 示威者, 表示, 示威, 影响 (events, New York, organization, city, reporters, people, demonstrators, said, protest, impact)	0.057/+	0.234/X
7	步步, 中国队, 转弯, 脑筋, 远去, 心计, 势力, 女友, 惊心, 控制 (step by step, the Chinese team, turning, brains, away, scheming, forces, girlfriend, startling, control)	0.052/+	0/X
8	华尔街, 苹果, 经济, 10, 可能, 第一, 发表, iPhone, 评论, 进行 (Wall Street, apple, economic, 10, maybe, first, published, iPhone, review, carried out)	0.109/X	0.090/X
9	日报, 乔布斯, 苹果, 创始人, 华尔街, 老爷, 蜡笔小新, 飞鹰, 幼儿园, 产品 (daily, Steve Jobs, Apple, founder, Wall Street, sir, Crayon, Eagle, nursery, products)	0.074/X	0/X
-	革命, 支持, 民主, 坦白, 美人, 发展, 希望, 推荐, 喜欢, 起来 (revolution, support, democracy, frankly, beauty, development, and hope, recommend, like, up)	positive	
-	危机, 鸿门宴, 贪婪, 消息, 失业, 解决, 呵呵, 抗议, 花心, 小子 (crisis, Banquet, greed, message, unemployment, solve, huh, protest, Fa, brat)	negative	
-	支持, 发展, 响应, 改革, 希望, 革命, 重要, 获得, 民主, 理解 (support, development, response, reform, hope, revolution, important, obtain, democracy, understanding)		positive
-	抗议, 游行, 危机, 不满, 贪婪, 爆发, 愤怒, 失业, 情绪, 解决 (protest, march, crisis, dissatisfaction, greed, outbreak, anger, unemployment, emotions, solve)		negative

The news media focused more on topic No.9 which discussed the social legal problem. News reports had the positive polarity towards topic No. 3. The positive topics of news and microblogs discussed the morality, i.e. people should help those ones who were injured in an accident.

4.5 Sentiment Evaluation

In order to evaluate the sentiment detection of our model, we introduce a method to convert topic-sentiment proportions into document sentiment predictions and compare the predictions with the sentiment labels $\{s_{cj}\}$ given in the MDS. Using our model, the positive proportion $p_{cj}(\text{pos.})$ and the negative proportion $p_{cj}(\text{neg.})$ of each document j in collection c can be calculated as:

$$p_{cj}(\text{pos./neg.}) \propto \sum_{k} \pi_{cj}(k) p_{cj}(z = k, l = \text{pos./neg.}) \tag{2}$$

where $\pi_{cj}(k)$ is the topic proportions and $p_{cj}(z = k, l)$ is the posterior topic-sentiment for the document. The sentiment distribution of a document is calculated as a summation on topics of the product of the topic proportion and

Table 6. Topics mined from "Little Yue Yue"

No.	Topics Mined in "Little Yue Yue"	Microblog	News
1	孩子, 小悦悦, 难道, 父母, 身边, 社会, 国人, 知道, 分享, 关注 (children, Little Yue Yue, do, parents, side, society, people, know, share, concerns)	0.054/X	0.047/X
2	小悦悦, 事件, 发生, 问题, 司机, 救人, 觉得, 父母, 女孩, 回复 (Little Yue Yue, incident, occurred, the problem, the driver, save, find, parents, girl, reply)	0.081/X	0.079/X
3	小悦悦, 蜡烛, 司机, 碾压, 走好, 视频, 广东, 民族, 第一, 博文 (Little Yue Yue, candles, drivers, rolling, take a good, video, Guangdong, ethnic, first, Bowen)	0.090/X	0.081/+
4	妈妈, 陈贤妹, 广州, 医院, 电话, 记者, 悦悦, 阿姨, 网友, 昨日 (mom, Chen Xian Mei, Guangzhou, hospitals, telephone, reporter, Yue Yue, aunts, friends, yesterday)	0.058/X	0/X
5	一路, 蜡烛, 走好, 世界, 良心, 别人, 离开, 车来车往, 蠟燭, 没有 (way, candles, take a good, world, conscience, others, leave, the car to drive to, candles, no)	0.051/X	0/X
6	大家, 看到, 责任, 中国, 已经, 良知, 真的, 一些, 所有, 评论 (we, see, responsibility, China, already, conscience, really, some, all, comments)	0.073/X	0.065/X
7	停止, 人们, 现在, 关心, 社会, 关怀, 路人, 没有, 中国人, 一下 (stop, people, now, care, social care, passers-by, no, the Chinese people, about)	0.077/X	0.058/X
8	蜡烛, 小悦悦, 反思, 今天, 保护, 新闻, 大家, 事件, 一直, 一起 (candles, Little Yue Yue, reflection, today, protecting, news, everyone, events, has, together)	0.063/X	0.048/X
9	法律, 社会, 出来, 事件, 10, 行为, 小悦悦, 认为, 美国, 感觉 (legal, social, out of, the event, 10, acts, Little Yue Yue, think, United States, feel)	0.065/X	0.100/X
-	道德, 天堂, 希望, 帮助, 见义勇为, 教育, 温暖, 其实, 相信, 起来 (morality, heaven, hope, help, courageous, education, warm, in fact, believe, up)	positive	
-	冷漠, 悲剧, 肇事, 可怜, 谴责, 见死不救, 冷血, 无情, 漠视, 悲哀 (indifference, tragedy, accident, poor, condemned, refused to help, cold-blooded, ruthless, disregard, sorrow)	negative	
-	道德, 见义勇为, 先生, 建设, 救助, 希望, 一定, 帮助, 精神, 重要 (moral, courageous, sir, construction, relief, hope, certainly, help, spiritual, important)		positive
-	肇事, 冷漠, 见死不救, 抢救, 谴责, 逃逸, 事故, 严重, 悲剧, 遭遇 (accident, apathy, refused to help, rescue, reprimand, escape, accident, serious, tragic, encounter)		negative

the topic-sentiment. The prediction of the document sentiment label is based on the ratio of p_{cj}(pos.) to p_{cj}(neg.). If the ratio is larger than the collection sentiment ratio calculated by the sentimental assignments, the document is positive, otherwise it is negative.

The accuracy results are shown in Table 7. In the "Game" domain the number of negative documents is far less than the number of positive documents and the accuracy of the negative in "Game" is low. Experiments also show that the performance of "Software" domain is not affected when modeling together. The model does not propagate the label-bias problem in one collection to the other collection.

Table 7. Sentiment evaluation: accuracy

sentiment	DVD	Electronics	Game	Software
positive	0.633	0.592	0.915	0.632
negative	0.613	0.639	0.205	0.612
overall	0.624	0.613	0.594	0.624

4.6 Comparison with JST Model

Both JST and our model extract topics with sentimental information; the JST in [10] find topics under positive, negative and neutral labels, and our model find topics with positive, negative proportions. We compare the results obtained from these two models based on the same MDS dataset.

Sentiment Analysis Comparison. Table 8 gives the accuracy comparison between our method, a lexicon method and the JST in [10]. Our result is better than the method based on the lexicon. The number of topics in JST is always a multiple of three labels. However, in our model, the number of topics is varied. According to [10], the number of topics under each label with the best classification is 15 on "DVD" but 1 on "Electronics". Our experimental result suggests the number of topics for both collections is 5. JST is designed for one collection and sets the number of topics for the sentiment classification while our work is proposed to choose the number of content topics in multi-collections and infer their sentiment polarities alongside.

Table 8. Sentiment analysis comparison (accuracy)

Methods	DVD	Electronics
Lexicon (shown in [10])	0.592	0.586
JST[10]	0.695 (K=15)	0.726 (K=1)
Ours	0.624 (K=5)	0.613 (K=5)

Topic Comparison. Our model can mine topics with different sentimental polarity and positive and negative topics for each collection. JST model mined each topic with the positive and negative topics. Their topic words are mixed with sentimental words. The result of both models are quite different. However we made the experiments on the same dataset and select some mined topics for comparison according to [10]. Tables 9, 10 are some results mined from DVD and Electronics domain. The first two lines are sentiment distributions found by our model, the third line is a topic found by our model with its sentiment polarity in brackets, and the fourth and fifth lines are the topics of JST from [10]. JST models the same number of topics under the positive, negative and neutral sentiments. Our model has mined not only different positive and negative topics for them, but also 5 topics with different sentimental polarity. The advantage of our model is that the number of topics is generated automatically.

5 Conclusions

In this paper, we present a topic-sentiment model for multiple text collections. A document is modeled as a mixture of components. Each component consists

Table 9. Topics mined from DVD

	Topic Words
Positive Topic (Ours)	like just good really even first great well love best
Negative Topic (Ours)	bad old hard trying boring used black understand wrong poor
Topic (Ours)	movie film see people think know story made life say [neutral]
Positive Topic (JST)	action good fight right scene chase hit art martial stunt
Negative Topic (JST)	horror scari bad evil dead blood monster zombi fear scare

Table 10. Topics mined from Electronics

	Topic Words
Positive Topic (Ours)	just sound like good great even quality well first really
Negative Topic (Ours)	used bad small hard old poor trying low expensive change
Topic (Ours)	product ipod problem player bought phone unit price headphones software [positive]
Positive Topic (JST)	mous hand logitech comfort scroll whell smooth feel accur track
Negative Topic (JST)	drive fail data complet lose failur recogn backup poorli error

of topics, positive and negative proportions and white noise. The HDP is used in the model for automatic selection of the number of components (topics) and reducing the dimension of parameter space used for topic-sentiment joint modeling. According to the experiments on MDS and EVENT corpus:

- The model can mine the topics and show different proportions of these topics in different collections. Some topics are shared among two collections, some topics are specific for each collection.
- The model can mine the positive and negative topics and identify the sentimental polarity towards those mined topics from each collection.

However, our model has assumed that multiple collections are discussing about the same event or the same type of topics. Each collection has only one positive and negative topic and many content topics. By setting different label proportion priors among topic words, sentimental words and white noise, we can get more fine-grained topics which is more suitable for the realistic world. The relationship between different priors and sentimental analysis will be explored in the future. We will do further experiments to analyze topics and sentiments from different media, different text collections about the same event.

References

1. Blei, D.: Probabilistic topic models. Communications of the ACM 55(4), 77–84 (2012)
2. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. the Journal of machine Learning research 3, 993–1022 (2003)

3. Blitzer, J., Dredze, M., Pereira, F.: Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. vol. 7, pp. 440–447. ACL, Prague, CZ (2007)
4. Cano, E., He, Y., Liu, K., Zhao, J.: A weakly-supervised bayesian model for violence detection from social media. In: The 6th International Joint Conference on Natural Language Processing (IJCNLP). ACL, Nagoya, Japan (October 2013)
5. Fang, Y., Si, L., Somasundaram, N., Yu, Z.: Mining contrastive opinions on political texts using cross-perspective topic model. In: Proceedings of the fifth ACM international conference on Web search and data mining. pp. 63–72. ACM, Seattle, Washington, USA (2012)
6. Friedman, H.H., Amoo, T.: Rating the rating scales. Journal of Marketing Management 9(3), 114–123 (1999)
7. Jo, Y., Oh, A.H.: Aspect and sentiment unification model for online review analysis. In: Proceedings of the fourth ACM international conference on web search and data mining. pp. 815–824. ACM, Hong Kong, China (2011)
8. Krippendorff, K.: Content Analysis: An Introduction to its Methodology. Sage Publications, Beverly Hills, CA, USA (1980)
9. Lin, C., He, Y.: Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM conference on Information and knowledge management. pp. 375–384. ACM, Hong Kong, China (2009)
10. Lin, C., He, Y., Everson, R., Ruger, S.: Weakly supervised joint sentiment-topic detection from text. IEEE Transactions on Knowledge and Data Engineering 24(6), 1134–1145 (2012)
11. Liu, B.: Sentiment analysis and subjectivity. In: Indurkhya, N., Damerau, F.J. (eds.) Handbook of Natural Language Processing, pp. 627–666. Chapman and Hall/CRC, Boca Raton, Florida, USA, 2 edn. (2010)
12. Mei, Q., Ling, X., Wondra, M., Su, H., Zhai, C.: Topic sentiment mixture: modeling facets and opinions in weblogs. In: Proceedings of the 16th international conference on World Wide Web. pp. 171–180. ACM, Banff, Alberta, Canada (2007)
13. Paul, M.J., Zhai, C., Girju, R.: Summarizing contrastive viewpoints in opinionated text. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp. 66–76. Association for Computational Linguistics, ACL, Massachusetts, USA (2010)
14. Teh, Y.W., Jordan, M.I.: Hierarchical Bayesian nonparametric models with applications, pp. 158–207. Cambridge University Press, Cambridge, UK (2010)
15. Teh, Y., Jordan, M., Beal, M., Blei, D.: Hierarchical dirichlet processes. Journal of the American Statistical Association 101(476), 1566–1581 (2006)
16. Xu, L., Lin, H., Pan, Y., Ren, H., Chen, J.: Constructing the affective lexicon ontology. Journal of the China Society for Scientific and Technical Information 27(2), 180–185 (2008), (In Chinese)