ENHANCED TWOFOLD-LDA MODEL FOR ASPECT DISCOVERY AND SENTIMENT CLASSIFICATION

NICOLA BURNS, YAXIN BI[*], HUI WANG & TERRY ANDERSON

School of Computing and Mathematics
University of Ulster. Newtownabbey, United Kingdom

Abstract: There is a need to automatically classify information from online reviews. Customers want to know useful information about different aspects of a product or service and also the sentiment expressed towards each aspect. This paper proposes an Enhanced Twofold-LDA model (Latent Dirichlet Allocation), in which one LDA is used for aspect assignment and another is used for sentiment classification, aiming to automatically determine aspect and sentiment. The enhanced model incorporates domain knowledge (i.e., seed words) to produce more focused topics and has the ability to handle two aspects in at the sentence level simultaneously. The experiment results show that the Enhanced Twofold-LDA model is able to produce topics more related to aspects in comparison to the state of arts method ASUM (Aspect and Sentiment Unification Model), whereas comparable with ASUM on sentiment classification performance. Additionally, an investigation is carried out to show the importance of research for customer satisfaction on various visual charts.

Keywords: sentiment analysis, aspect discovery, sentiment classification, topic modeling

1. Introduction

The Internet is a common way of life for most people these days with more and more tasks being carried out online. For example, if a customer buys a product online, they can leave a review indicating how they felt about the product. Reviews are used for many different domains, such as products, movies and holidays. Reviews are continuing to be a popular way of expressing views on products or services and also a means of seeking information about other reviewer opinions. This has resulted in a growth in the amount of reviews, which are often only labeled with an overall rating.

Reviews contain much more useful information than an overall rating, including opinions expressed towards different aspects of the product or service. A movie review could have an overall rating of 5 stars, but the reviewer could still express negative sentiment towards one of the aspects. Reading all the reviews available would be very time consuming and difficult to make comparisons. There is therefore a need to automatically classify reviews, indicating aspects and sentiment in a way which is easy to perceive and compare.

To this end, we have developed the Twofold-LDA model, it assigns each sentence to an aspect, quantifies the results of positive and negative sentiment of each aspect and visualize them in an intuitive way [1]. The motivation of the Twofold-LDA model was to add supervision to the model in the form of domain knowledge to direct the focus of topics towards more relevant aspects than those produced by the standard LDA model

[*] Corresponding author, email: y.bi@ulster.ac.uk

[2]. The Twofold-LDA model also incorporates multiple aspects in one sentence removing the one aspect, one sentence assumption which research has previously used. The results produced by the Twofold-LDA model are then transformed into findings from which a customer or manufacturer can benefit in terms of visual charts that are useful, easy to read, easy to compare and quick to understand.

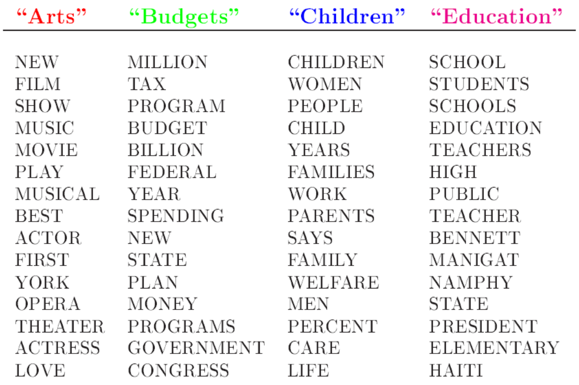
The Twofold-LDA model is done in two separate stages, discovering the aspect related to each sentence and discovering the sentiment of the same sentence. Meanwhile we incorporate part-of-speech tagging (POS) into the Twofold-LDA modeling process, whereby improving its sentiment classification performance, in this way we call the Twofold-LDA model as an Enhanced Twofold-LDA model. We compare the Enhanced Twofold-LDA model with the Twofold-LDA model to show how efficiency has improved. We also compare the Enhanced Twofold-LDA model with ASUM and find that the Enhanced Twofold-LDA model performs better for aspect discovery, however ASUM performs better for sentiment classification.

The paper is outlined as follows; in Section 2, we first present the Enhanced Twofold-LDA model. We then present experimental results in Section 3, including aspect discovery, sentiment classification and the comparison of the Enhanced Twofold-LDA model and ASUM. Section 4 discusses related work and finally, Section 5 gives a conclusion of the work.

2. Enhanced Twofold-LDA model

Figure 1 shows an example of the standard LDA model output, usually 30 to 100 topics manually labelled with the aspects that relate to the words in each topic. This manual work of making sense of the output can be quite time-consuming and would not be easily understood by an end user.

Previous research using natural language processing techniques shows information in graphical form which is much more user friendly. Figure 2 shows an example of a chart comparing two digital cameras. From the chart we can clearly see that Digital Camera 1 is better in nearly every aspect and both cameras have equal positive and negative opinion on their weight. This chart means that customers no longer need to read through



"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Fig. 1. Standard LDA output [2]

countless reviews to find the useful information they require. Also, the chart allows a customer to compare 2 products along side each other.

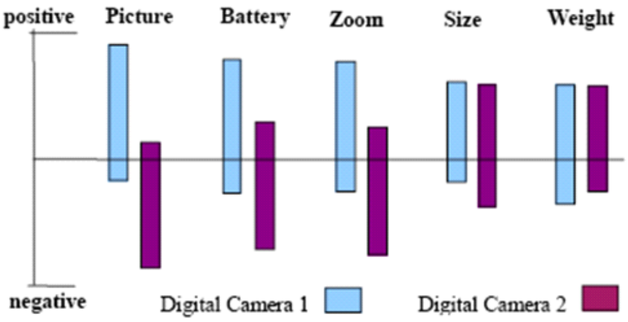
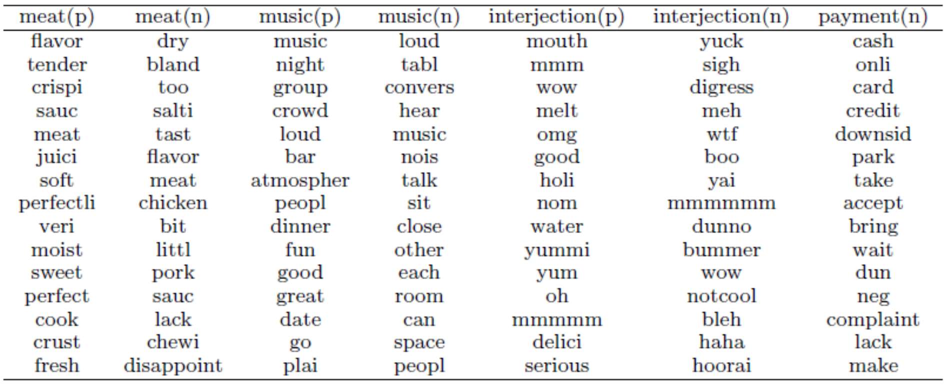


Fig. 2: Visual comparison of two digital cameras [3]

Figure 3 shows an example of the results outputted from ASUM. This is a unified model which uses the LDA model as a basis for ASUM [4]. This output requires manual labeling for each topic e.g., music (*n*). This would take a lot of time as there could be 30 or 100 topics and the findings are still difficult to comprehend. A customer or manufacturer would not find this easy to read.

In this paper, we propose an Enhanced Twofold-LDA model, which is aimed to:

- improve the efficiency of the Twofold-LDA model in automatically producing visual charts.
- improve sentiment classification performance by incorporating POS tags into the sampling process.
- compare the proposed Enhanced Twofold-LDA model with a recent comparable study, namely ASUM.



meat(p)	meat(n)	music(p)	music(n)	interjection(p)	interjection(n)	payment(n)
flavor	dry	music	loud	mouth	yuck	cash
tender	bland	night	tabl	mmm	sigh	onli
crispi	too	group	convers	wow	digress	card
sauc	salti	crowd	hear	melt	meh	credit
meat	tast	loud	music	omg	wtf	downsid
juici	flavor	bar	nois	good	boo	park
soft	meat	atmospher	talk	holi	yai	take
perfectli	chicken	peopl	sit	nom	mmmmmm	accept
veri	bit	dinner	close	water	dunno	bring
moist	littl	fun	other	yummi	bummer	wait
sweet	pork	good	each	yum	wow	dun
perfect	sauc	great	room	oh	notcool	neg
cook	lack	date	can	mmmmm	bleh	complaint
crust	chewi	go	space	delici	haha	lack
fresh	disappoint	plai	peopl	serious	hoorai	make

Fig. 3: Restaurant review senti-aspect [4]

2.1. *Model Description*

This section describes the Twofold-LDA model, including how seed words are integrated into this model. It also contains a description how the model is improved to develop the Enhanced Twofold-LDA model which includes incorporating POS tags and automatically producing visual charts.

2.1.1. LDA model

Firstly, we will look at the original LDA shown in Figure 4 [2].

In this figure, $\boldsymbol{w}$ represents the words in the document $d \in D$ (a set of documents) and z signifies the hidden aspect from which the word $\boldsymbol{w}$ is generated. Priors prevent the model from over fitting; β is the Dirichlet prior vector for $\varphi \in T$ (a list of topics) and α is the Dirichlet prior vector for θ; φ is a multinomial distribution over words and θ is a multinomial distribution over aspect. Each document is generated by choosing a distribution over aspects θ, which determines *P*(z) for words in that document. Each word is then generated randomly from an aspect j using this distribution, then picking a word from that aspect according to $p(w|z = j)$. For document d, $\varphi_j^{(w)} = p(w|z = j)$ and $\theta_j^{(d)} = p(z = j)$ (N_d is a number of words in *d*). LDA assumes the following generative process:

- For each document d, choose a distribution $\theta_d \sim Dirichlet\ (\alpha)$
- For each word w_i in the document d
  - Choose aspect z_i ~ Multinomial (θ_d)
  - Generate word w_i from $p(w_i|z_i, \beta)$, a multinomial probability conditioned on the aspect z_i

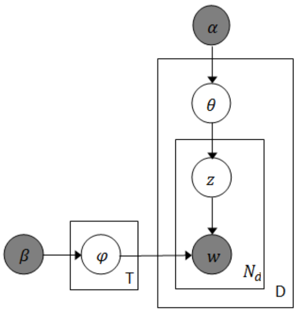


Fig 4. LDA plate diagram

2.1.2. The Twofold-LDA model

Figure 5 shows the Twofold-LDA model, here we can see that the two LDAs run separately on the same dataset. One LDA model is used for aspect extraction and one LDA model is used for sentiment classification. We input seed words into the model for both aspects and sentiment, the light grey nodes highlight our semi-supervision feature. The generative process is as follows:

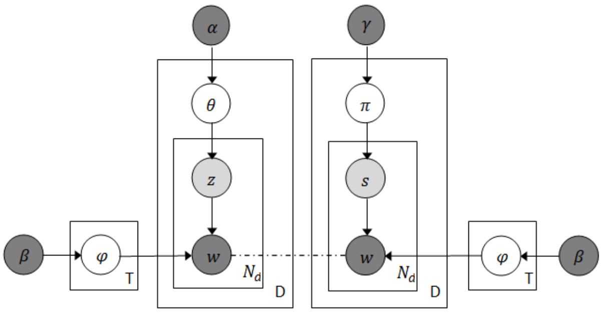


Fig. 5. Plate diagram of Twofold-LDA

For aspect assignment:

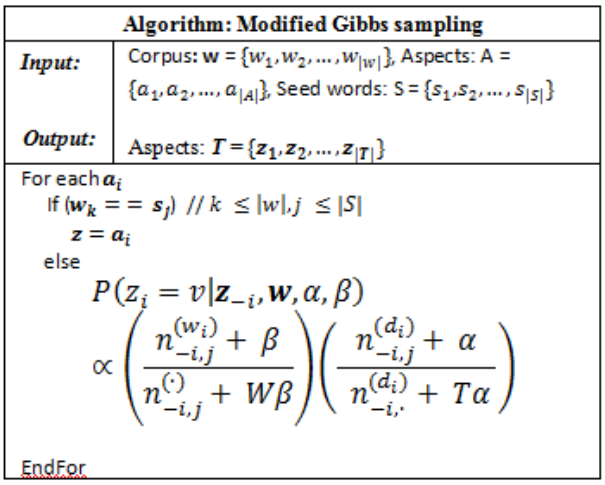
- For each document d, choose a distribution $\theta_d \sim Dirichlet\ (\alpha)$
- For each word w_i in the document d
  - Choose aspect z_i ~ Multinomial (θ_d)
  - Generate word w_i from $p(w_i|z_i, \beta)$, a multinomial probability conditioned on the aspect z_i

For polarity assignment:

- For each document d, choose a distribution $\pi_d \sim Dirichlet\ (\gamma)$
- For each word w_i in the document d
  - Choose sentiment s_i ~ Multinomial (π_d)
  - Generate word w_i from $p(w_i|s_i, \beta)$, a multinomial probablility conditioned on the sentiment s_i

To incorporate the seed words into the Twofold-LDA model we took a similar approach in [5] and revised the Gibbs sampling equation. This was done at two steps, first at initialization we set all the seed words to the relating aspect. For example, seed words {*price*, *bought*, *buy*, etc.} may all be set to Aspect 4 which might correspond to the aspect *Value*. Secondly, the Gibbs sampling equation is modified to keep these aspect words in the correct aspect, i.e., sampling is not performed on aspect words. Keeping words such as *price*, *bought* and *buy* in the same aspect will encourage relevant words to be classified into this aspect. The modified Gibbs sampling equation can be seen in Figure 6. If a word belongs in one of our seed sets (there is a list of seed words for each aspect), it is assigned to the correct aspect, otherwise the Gibbs sampling is performed.

In [6], the authors use Markov chains, specifically Collapsed Gibbs Sampling to discover hidden aspects $\boldsymbol{z}$ given observed words $\boldsymbol{w}$. To sample individual z_i, we use the full conditional probability where $n_{-i,j}^{(w_i)}$ is the number of times, word w_i is generated by aspect j, and $n_{-i,j}^{(d_i)}$ is the number of times, aspect j is used in document d. $n_{-i,j}^{(\cdot)}$ indicates that the counts are taken excluding the value of z_i. *W* consists of unique words which are generated from $\boldsymbol{w}$ and *T* is a collection of aspects. The first ratio indicates the probability of w_i under aspect j, and the second ratio indicates the probability of aspect j in document d [6].



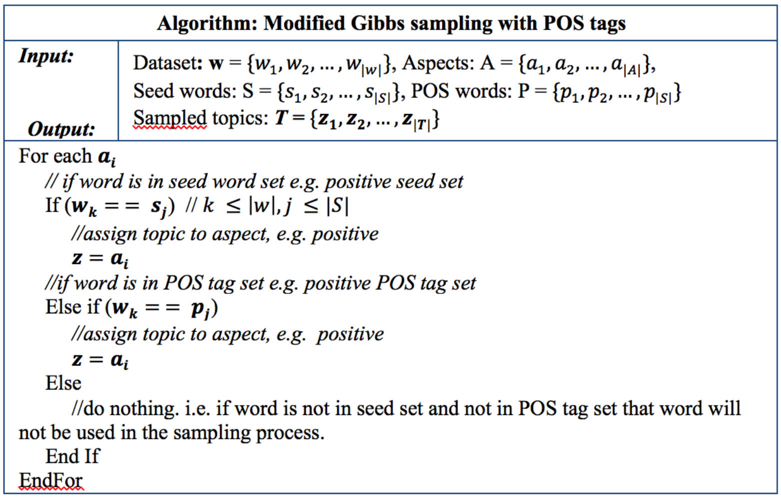
```
Algorithm: Modified Gibbs sampling
Input:   Corpus: w = {$w_1, w_2, \dots, w_{|w|}$}, Aspects: A =
         {$a_1, a_2, \dots, a_{|A|}$}, Seed words: S = {$s_1, s_2, \dots, s_{|S|}$}
Output:  Aspects: $\boldsymbol{T} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \dots, \boldsymbol{z}_{|T|}\}$
For each $\boldsymbol{a}_i$
  If ($\boldsymbol{w}_k == \boldsymbol{s}_j$) // $k \leq |w|, j \leq |S|$
    $\boldsymbol{z} = \boldsymbol{a}_i$
  else
    $P(z_i = v|\mathbf{z}_{-i}, \mathbf{w}, \alpha, \beta) \propto \left(\frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta}\right)\left(\frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}\right)$
EndFor
```

Fig. 6. Modified Gibbs sampling equation adapted from Griffiths and Steyvers (2004) [6]

2.1.3. *The Enhanced Twofold-LDA model*

From the literature it can be seen that within the natural language processing domain, using POS tagging is a common and effective way to achieve high sentiment classification accuracy. We therefore incorporate POS tagging into our Twofold-LDA model as domain knowledge, resulting in an Enhanced Twofold-LDA model, to the best of our knowledge this has not been done so far. To do this we again modify the Gibbs sampling equation, shown in Figure 7. First the Enhanced Twofold-LDA model identifies the POS tags separately from the modeling process. This is done using the Standford POS tagger which identifies the POS tag for each word in the dataset [47]. A matrix is then created, which contains each word and its associated tag with the highest probability. For example, a word in the TV dataset will be classified as a noun in the matrix if that word is more likely to be a noun than a verb. The matrix, $P = \{p_1, p_2, \ldots p_{|p|}\}$, is then used as input to the modified Gibbs sampling algorithm as described in Figure 7.

With the modified Gibbs sampling algorithm, for each aspect $A = \{a_1, a, \ldots a_{|A|}\}$, e.g., positive and negative, if a word w belongs to one of the words in the aspect seed sets $S = \{s_1, s_2, \ldots s_{|S|}\}$, the word w is assigned to that particular topic $T = \{z_1, z, \ldots z_{|T|}\}$. If the word w belongs to one of the words in the POS set $P = \{p_1, p, \ldots p_{|P|}\}$, the word is assigned to that topic $T = \{z_1, z, \ldots z_{|T|}\}$ (e.g., positive may be Topic 1 and negative may be Topic 2). As a result, the output of the sampling process is a list of topics that have been sampled using aspect seed words in conjunction with words with relevant POS tags. When the modified Gibbs sampling with POS tags is used, only seed words or words labeled with particular POS tags influence sampling and the resulting aspects.



Algorithm: Modified Gibbs sampling with POS tags

Input:	Dataset: $\mathbf{w} = \{w_1, w_2, \ldots, w_{|w|}\}$, Aspects: $A = \{a_1, a_2, \ldots, a_{|A|}\}$, Seed words: $S = \{s_1, s_2, \ldots, s_{|S|}\}$, POS words: $P = \{p_1, p_2, \ldots, p_{|S|}\}$
Output:	Sampled topics: $\boldsymbol{T} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_{|T|}\}$

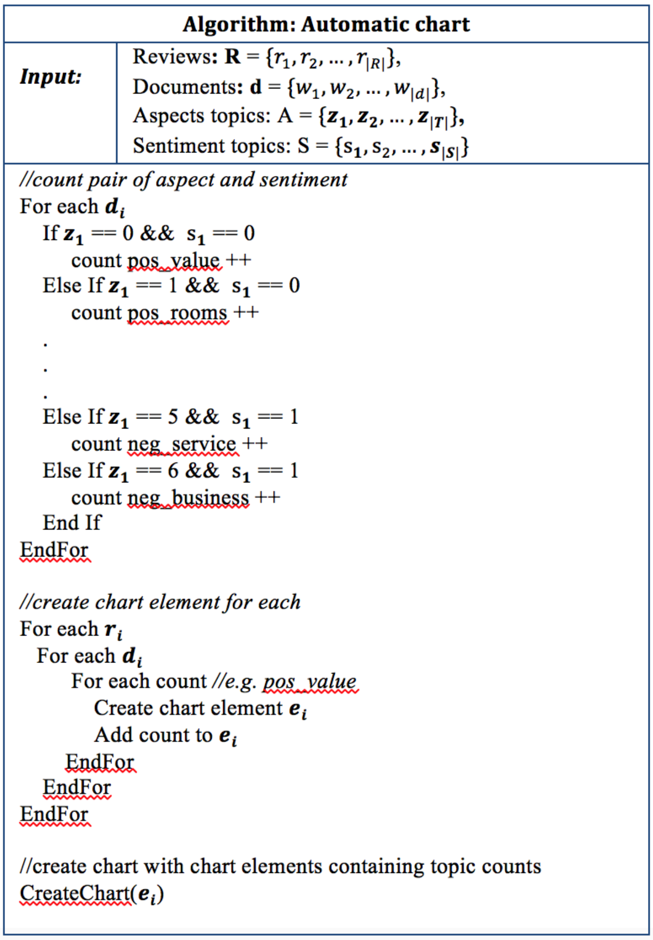
```
For each $\boldsymbol{a}_i$
    // if word is in seed word set e.g. positive seed set
    If ($\boldsymbol{w}_k == \boldsymbol{s}_j$) // $k \le |w|, j \le |S|$
        //assign topic to aspect, e.g. positive
        $\boldsymbol{z} = \boldsymbol{a}_i$
    //if word is in POS tag set e.g. positive POS tag set
    Else if ($\boldsymbol{w}_k == \boldsymbol{p}_j$)
        //assign topic to aspect, e.g.  positive
        $\boldsymbol{z} = \boldsymbol{a}_i$
    Else
        //do nothing. i.e. if word is not in seed set and not in POS tag set that word will
not be used in the sampling process.
    End If
EndFor
```

Fig. 7. Modified Gibbs sampling equation including POS tags adapted from [6]

The Enhanced Twofold-LDA model also includes the following algorithm to automatically visualize graphical charts, which consists of three main steps as illustrated Figure 8:

1. First for each document d, the algorithm counts the number of pairs of aspect z_1 and sentiment s_1 for each sentence. For example, sentence 0 is identified as *Value* and *Positive*.

2. For each review R, a chart element is created for each of the counts, e.g., pos_value.
3. Finally, the chart elements which contain the counts are used to create the charts.



Algorithm: Automatic chart

Input:	Reviews: $\mathbf{R} = \{r_1, r_2, \dots, r_{\vert R\vert}\}$, Documents: $\mathbf{d} = \{w_1, w_2, \dots, w_{\vert d\vert}\}$, Aspects topics: A = $\{\boldsymbol{z_1}, \boldsymbol{z_2}, \dots, \boldsymbol{z_{\vert T\vert}}\}$, Sentiment topics: S = $\{\mathrm{s_1}, \mathrm{s_2}, \dots, \boldsymbol{s_{\vert S\vert}}\}$

```
//count pair of aspect and sentiment
For each $\boldsymbol{d_i}$
   If $\boldsymbol{z_1}$ == 0 && $\mathrm{s_1}$ == 0
       count pos_value ++
   Else If $\boldsymbol{z_1}$ == 1 && $\mathrm{s_1}$ == 0
       count pos_rooms ++
   .
   .
   .
   Else If $\boldsymbol{z_1}$ == 5 && $\mathrm{s_1}$ == 1
       count neg_service ++
   Else If $\boldsymbol{z_1}$ == 6 && $\mathrm{s_1}$ == 1
       count neg_business ++
   End If
EndFor

//create chart element for each
For each $\boldsymbol{r_i}$
  For each $\boldsymbol{d_i}$
      For each count //e.g. pos_value
         Create chart element $\boldsymbol{e_i}$
         Add count to $\boldsymbol{e_i}$
     EndFor
   EndFor
EndFor

//create chart with chart elements containing topic counts
CreateChart($\boldsymbol{e_i}$)
```

Fig. 8: Automatic chart algorithm.

3. Experimental Results

We now report the experimental results to show the advantage of the Enhanced Twofold-LDA model for aspect discovery and sentiment classification. For aspect discovery, we measure performance using precision, recall and f-measure, respectively [48]. For sentiment classification we calculate the accuracy of each sentence, then calculate the overall accuracy for each review and compare this against the benchmark techniques. For experiments we use a 10-fold cross validation. Symmetric hyperparameters α = 0.1 and β

= 0.01 were used and Markov chains run for 1000 samples after the initial assessment [7].

3.1. *Datasets*

We use 3 datasets, 2 of which are publically available, namely, TripAdvisor and Mp3 datasets. Then we have our manually obtained dataset, i.e., the TV review dataset. A summary of the properties of these datasets can be found in Table 1

Table 1. Properties of datasets.

	TripAdvisor	TV (all sentences)	TV (aspect sentences)	Mp3
# reviews	1,850 (108.891)	14,450	13,923	23,945
# sentences	1,513,655	158,645	81,919	82,823
# pruned words	10,961,805	926,382	574,676	664,989
Avg. # sentences in review	13.90	10.99	5.88	3.46
Avg. # pruned words in review	100.67	64.11	41.28	27.77

TripAdvisor dataset. This dataset is made up of hotel reviews from the TripAdvisor website. It has been used by another work to jointly detect aspect and sentiment [7]. This is the largest one of the 3 datasets, containing 108,891 reviews. The reviews often contain long sentences of free format text with a five star reading. We classify reviews with 4 or 5 stars as positive and 1 or 2 stars as negative. We discard reviews with a 3 star rating as these are considered neutral. Reviewers rate the following aspects; *value, room, location, cleanliness, check in/front desk, service* and *business service*. We choose this dataset as our model requires seed words on each aspect to incorporate into the model, seed words have already been extracted via a bootstapping algorithm and made available for this particular dataset [7], which can be particularly suited for our experiments.

Mp3 dataset. The Mp3 dataset is the 2nd publically available dataset consisting of Mp3 reviews from Amazon. The Mp3 dataset tends to have shorter sentences, which can have an impact on the model [8]. We again discard neutral reviews and use 4 or 5 stars as positive and 1 or 2 stars as negative. This dataset does not have any seed words available, nevertheless we used the aspects that reviewers comment on from reevoo.com as the main aspects for our model, these aspects include *design, sound, value* and *ease of use.*

TV dataset. Finally, the TV dataset was manually obtained. We extracted reviews from Amazon in September 2010. For our experiments we again discard the neutral reviews, leaving only the positive and negative reviews. This dataset is the smallest of these three datasets , it contains 12,374 positive reviews and 2,076 negative reviews. We also have available seed words for the TV dataset which were obtained in a previous study [1]. The TV aspects include *design, image quality, sound quality, value* and *ease of use*.

We applied the same process as in [7] for the TripAdvisor, TV (aspect sentences) and Mp3 datasets and discard all sentences that do not contain any aspect seed words, as not all reviews contain all the aspects. The TripAdvisor datasets then takes all the reviews for each hotel and create one review, so there are 1,850 hotels and therefore 1,850 reviews [7, 9]. The format of the TV dataset downloaded does not allow us to concatenate the reviews for each TV so we keep these reviews separate. The mp3 dataset contains only one mp3, thus we also keep these reviews separate.

3.2. *Aspect extraction*

Our previous work proposed the Twofold-LDA model and made a comparison to the standard LDA model. This work therefore makes a comparison of the Twofold-LDA model and a comparable model, namely ASUM. We then apply the Enhanced Twofold-LDA model to aspect discovery which results in a decrease in performance.

3.2.1. Comparison of the Twofold-LDA model and ASUM

A popular method in the recent research has been to jointly model aspect and sentiment. Joint models firstly have the benefits of efficiency, and secondly they contain common knowledge between aspects and sentiment. The literature shows, however, that the results produced by joint models are not user friendly. An example is the output of ASUM, the output is a simple list of words for each topic in order of probabilities as shown in Figure 3 [4]. Converting these results into a chart would prove difficult as there are a large number of topics produced which must all be hand labeled. ASUM is an extended LDA model which models aspect and sentiment together to produce sentiments towards different aspects using senti-aspects (pairs of {aspect, sentiment}).

Table 2. Comparison of ASUM and the Twofold-LDA model.

	ASUM	**Twofold-LDA**
Identify aspects	✓	✓
Identify sentiment	✓	✓
Jointly identify aspects and sentiment	✓	
Easily convert results into chart		✓
Incorporate sentiment seed words	✓	✓
Incorporate aspect seed words		✓
Sentence-level modeling	✓	✓
Manual work to determine results	✓	
Model requires user rating		
Identify multiple aspects in one sentence		✓

Includes negation.	✓

Table 2 shows a comparison of ASUM and the Twofold-LDA model clearly showing the similarities and differences. The main differences between the two models are that ASUM jointly models aspect and sentiment, whereas the Twofold-LDA model models aspect and sentiment separately and combines their results. Also, the Twofold-LDA model requires additional aspect seed words prior to the modeling process which ASUM does not require. However, this results in no manual work for the Twofold-LDA model when the results are produced as is required for ASUM.

To investigate how well aspects are discovered we compare the results of the Twofold-LDA model and ASUM. The Twofold-LDA model is setup as follows: hyperparameters are set to $\alpha = 0.1$ and $\beta = 0.01$ and Markov chains run for 1000 samples as previously suggested in ASUM [7]. The three datasets used for experiments are: TripAdvisor, TV (aspect sentences) and TV (all sentences). For the Twofold-LDA model, the number of topics is set to 7, one topic for each of the 7 TripAdvisor aspects. For ASUM the number of topics is set to 30, these topics were then manually labeled with aspects. The cosine similarity measure is used to compare the top words in these topics [10]. Experiments were first carried out with ASUM topics set to 7, however ASUM was unable to discover all aspects with the topics set this low. Also, 4 of the 7 negative topics produced were related to the *Room* aspect. Therefore, the number of topics was increased to make a fairer comparison.

To make a comparison, we chose one topic from each dataset, *Room* (TripAdvisor), *Sound quality* (TV – aspect sentences) and *Image quality* (TV – all sentences). The seed words are color coded to indicate the aspects discovered. Note that the Twofold-LDA model uses aspect and sentiment seed words, whereas ASUM uses only sentiment seed words in their respective modeling process. We use the top 50 words of each aspect discovered by the two models for comparison.

Table 3a shows the *Room* topic discovered by the Twofold-LDA model. It is clear that the contents of this topic relate to rooms, 94% of the topic words are aspect seed words and the remaining words are also related to rooms. Table 3b shows the positive and negative *Room* topics produced by ASUM, these topics were manually labeled as *Room*. Note, only aspects manually labeled as sentiment is discovered automatically. The colored words in Table 3b indicate that the topics contain words from multiple aspects. Each of the topics produced by ASUM contained words which may relate to another aspect. The positive *Room* topic shows many positive words e.g., nice, amazing and fantastic. The negative *Room* topic reveals neither positive nor negative words in the top 50 words discovered. Additional experiments for sentiment classification are carried out later to investigate sentiment results and indicate the TripAdvisor dataset may perform poorly for identifying negative sentiment.

Table 3a: Twofold-LDA *Room* (TripAdvisor) topic

Room	**room**, **stay**, **night**, **bed**, **view**, **floor**, **bathroom**, **comfortable**, **shower**, **quiet**, **suite**, **noise**, **size**, **spaciou**, **window**, **modern**, **square**, **air**, **read**, **tv**, chair, **double**, **balcony**, **sleep**, **space**, **bath**, **light**, **king**, **bedroom**, **decor**, **tower**, **house**, **upgrade**, **pillow**, housekeeping, **courtyard**, **inside**, **separate**, **furniture**, **queen**, **sink**, **carpet**, **apartment**, **renovation**, **channel**, **condition**, **soap**, conditioner, **bathtub**, **mirror**,

Table 3b: ASUM *Rooms* (TripAdvisor) positive and negative topics

Room (Positive)	**room, view, floor,** great, ocean, hotel, very, **suite,** nice, **stayed, balcony,** city, **tower,** **upgraded, quiet,** beautiful, **window, night,** building, **overlooking,** you, **clean,** my, corner, got, us, **pool,** top, large, **beach,** booked, good, **bed,** lovely, spaciou, amazing, fantastic, harbour, facing, street, **front, location,** wonderful, side, bay, deluxe, partial, given, river, **comfortable**
Room (Negative)	**room, bed,** us, **floor**, booked, hotel, **suite, king,** my, told, **night, double, view,** asked, requested, **upgraded, upgrade,** got, available, given, me, did, day, **desk, arrived,** first, another, **front,** request, **smoking, stay**, **reservation**, next, moved, get, **non-smoking**, upon, arrival, said, **checked, size**, even, extra, ocean, **check**, queen, gave, paid, offered, **club**

Table 4a shows the *Sound quality* topic produced by the Twofold-LDA model. This topic contains 22% seed words. Only *Sound quality* seed words are contained in the topic. Table 4b shows each of the positive and negative topics discovered by ASUM which are manually labeled as *Sound quality*. There is a small amount of overlap with the ASUM topics, the positive topic contains only one word from another aspect and the negative topic contains three words from other aspects. The positive *Sound quality* topic contains a number of positive words, e.g., good, great and better. The negative *Sound quality* topic contains negative words, e.g., problem, complaint and bad.

Table 4a: Twofold-LDA *Sound quality* (TV – aspect sentences) topic

Sound quality	**sound**, tv, **speaker**, not, **audio**, good, quality, system, **volume**, great, **surround**, use, home, set, theater, don, **hear**, **loud**, fine, **noise**, issue, receiver, output, room, review, problem, external, low, **auto**, **stereo**, little, bad, **music**, pretty, ok, am, people, decent, hooked, level, buzzing, watching, adequate, isn, built, complaint, poor, tinny, hook, expect

Table 4b: ASUM *Sound quality* (TV – aspect sentences) positive and negative topic

Sound quality (Positive)	**sound, speaker,** tv, system, you, **surround**, good, quality, my, great, home, theatre, **audio**, use, **picture**, better, set, very, external, get, don, really, need, just, want, receiver, fine, **volume,** **stereo, buy**, hooked, do, me, am, pretty, even, built, through, much, **using**, ok, small, hook, what, your, adequate, best, going, mov, room
Sound quality (Negative)	**sound,** tv, **volume, speaker,** my, you, **loud, hear,** system, **surround, audio**, quality, low, don, level, even, tinny, very, problem, set, enough, just, fine, complaint, much, use, bad, small, little, review, turn, issue, bit, **setting**, really, ba, room, too, me, better, clear, home, people, theatre, what, bedroom, adequate, need, isn, do

Finally, Table 5a shows the *Image quality* topic produced by the Twofold-LDA model. This topic contains 32% seed words. Table 5b shows the top 50 words in each of the positive and negative topics discovered by ASUM and manually labeled as *Image quality*. Again, based on the seed word colors it is clear that each topic produced by ASUM contains words from other aspects. The positive *Image quality* topic contains a number of positive words, e.g., better, good and great. The negative *Image quality* topic contains no negative words and one positive word, i.e., happier, in the top 50 words discovered.

Table 5a: Twofold-LDA *Image Quality* (TV – all sentences) topic

Image quality	**picture**, tv, **screen**, **lcd**, quality, **hd**, **plasma**, not, hdmi, black, samsung, hdtv, great, good, **led**, set, **viewing**, **light**, **image**, **sharp**, sony, **digital**, 1080p, room, **contrast**, **bright**, best, **monitor**, difference, **signal**, dark, excellent, level, amazing, angle, **motion**, panasonic, tvs, model, seen, ve, inch, watching, 120hz, brightne, issue, 720p, scene, little, don

Table 5b: ASUM *Image Quality* (TV – all sentences) positive and negative topic

Image quality (Positive)	tv, **picture**, **lcd**, **plasma**, samsung, quality, better, my, **led**, black, set, good, you, sony, much, **price**, **color**, **model**, **screen**, great, very, best, look, panasonic, just, even, tvs, **contrast**, difference, year, level, **sharp**, compared, **feature**, **viewing**, really, **hdtv**, 1080p, lg, **size**, me, ser, what, still, think seen, **sound**, new, **image**, angle
Image quality (Negative)	**hdtv**, **lcd**, 1080p, samsung, sony, bravia, **hz**, tv, **led**, black, my, panasonic, 720p, **plasma**, 46-inch, ser, 120hz, **purchased**, **color**, 32-inch, lg, touch, amazon, **sharp**, red, xbr, 240hz, viera, second, inch, 52-inch, 40-inch, 55-inch, **bought**, 42-inch, first, vizio, **model**, vs, 37-inch, toshida, v-ser, ordered, aquo, In46a650, **hd**, happier, rate, own, returned

The *rand index* is then used to compare the aspects discovered by the Twofold-LDA model and ASUM for each dataset namely [47]; TripAdvisor, TV (aspect sentences) and TV (all sentences). To make the comparison, the seed words discovered by each topic for the Twofold-LDA model were compared against the seed words discovered for each topic for ASUM.

Table 6: Rand index of the Twofold-LDA model and ASUM.

	TripAdvisor	**TV (aspect sentences)**	**TV (all sentences)**
Twofold-LDA model	1	1	1
ASUM	0.77 (pos) 0.76 (neg)	0.86 (pos) 0.75 (neg)	0.78 (pos) 0.81 (neg)

Table 6 shows that the Twofold-LDA model has a rand index score of 1 for each dataset, which shows that the sets (clusters) of manually labeled aspect seed words and the sets of aspect-based keywords in the topics produced by the Twofold-LDA model are the same. The highest score of 1 indicates that there are overlapping words in each of the aspects with the clusters of the labeled aspect seed words. ASUM reaches reasonably high scores for the positive and negative topics for each of the three datasets indicating that the clusters of seed words and clusters of aspect-based keywords in the topics discovered by ASUM are fairly similar. These results show that the Twofold-LDA model produces topics which are more related to the aspect seed words than those produced by ASUM. Also, the Twofold-LDA model contains no overlapping between the aspects discovered.

3.2.2. Aspect discovery using POS tags

For an additional experiment for aspect discovery, we decided to apply the modified Gibbs Sampling equation with POS tags described in Figure 7, and apply it to the aspect model. The equation described in Figure 7 is aimed at sentiment rather than aspects. As mentioned in the related work POS tagging has been used for association rule mining to extract candidate aspects, this is done using noun phrases [14, 15]. The previous studies have stated that aspects are usually nouns [16]. We therefore use a set of words with noun tags as input into the modified Gibbs Sampling equation shown in Figure 7. As a result, only noun words can be assigned to an aspect topic.

Table 7. Comparison of the Enhanced Twofold-LDA model with and without POS tags.

	Design	*Sound*	*Value*	*Ease*
Without POS tags	84.262	85.634	85.692	74.923
With POS tags	81.624	83.821	83.720	76.031

Table 7 shows the f-measure results of each aspect in the Mp3 dataset. We can see that in 3 out of 4 aspects, incorporating POS tags actually decreases the performance of the Enhanced Twofold-LDA model. These results provide evidence that incorporating POS tagging into a topic model when performing aspect discovery can have a negative effect on performance. A reason for this decrease in performance would be that the modified Gibbs sampling equation with POS tagging only uses nouns and aspect seed words in its sampling process, this reduces the coverage of the sampled words, thereby reducing the performance of aspect discovery. In the following section, we carry out the same experiment on all 3 datasets for sentiment classification and achieve improved performance in almost every case.

3.3. *Sentiment classification*

In this section, we quantitatively evaluate the sentiment side of the Enhanced Twofold-LDA model. To determine sentiment, we use the same approach as aspect discovery, we incorporate sentiment seed words into the Enhanced Twofold-LDA model via the modified Gibbs sampling equation shown in Figure 6, we also incorporate POS tagging into the Enhanced Twofold-LDA model via the modified Gibbs sampling equation with POS tags shown in Figure 7. We then compare the Enhanced Twofold-LDA model to the state-of-art classification techniques, namely Naive Bayes and language model (a model of probability distribution of generating each word) [9]. Also, a comparison is made against a recent comparable study, ASUM.

3.3.1. Experimental Setup

Previous research has stated that sentiment classification is more difficult than aspect based classification [17, 18, 20]. This is because aspects can be recognized by the co-occurrence of keywords whereas, sentiment can be expressed more subtly. We made an observation that sentiment classification techniques which incorporate POS tagging can achieve good performance for classification, possible because this approach takes the structure and context of a sentence into consideration [14, 15, 21]. The Enhanced Twofold-LDA model is based on the LDA model which is usually used for identifying aspects. To discover sentiment, we decided to explore how we can use sentiment techniques i.e., incorporating POS tags, and incorporate them into the Enhanced Twofold-LDA model. We therefore investigate using different sentiment prior information then extend the model to incorporate POS tags. We first classify each sentence then use these results to classify each review, i.e., if a review is made up of 5 positive sentences and 2 negative sentences, the overall review will be considered positive. We classify at sentence-level and convert into review-level as we are using the overall star rating as ground truth. Some positive reviews may contain negative sentences or vice versa, thus we do not make an assumption that a positive review will have all positive sentences.

We evaluate with various types of prior information. Table 4 shows the amount of positive and negative seed words for each type of prior information. The types are as follows:

- Paradigm – a list of general sentiment words taken from Jo and Oh (2011) [4].
- Paradigm++ – this is Paradigm with the addition of sentiment seed words from the TripAdvisor dataset.
- Subjectivity lexicon – this is a large general list of sentiment words. We use only the strongly subjective words from the list [17].

As the Enhanced Twofold-LDA model uses a bag-of-words approach if a reviewer says ‘not good’, the word ‘good’ will highly likely be classified as positive. Therefore, we add negation to the Enhanced Twofold-LDA model by changing the sentiment if the word ‘not’ comes before a positive word. Therefore, the word ‘nice’ in the sentence ‘the staff were really nice’ will be classified positive and the word ‘nice’ in the sentence, ‘the staff were not nice’ will be classified negative.

To incorporate POS tags we used the Stanford POS tagger. We use a matrix for the most common tag for each word. It has been stated that opinion words are usually adjectives, verbs or adverbs [16]. We thus made a set of words with opinion tags to be incorporated into the Enhanced Twofold-LDA model i.e., all words from the dataset which are adjectives, verbs or adverbs. We then input the opinion set into our modified Gibbs sampling equation in Figure 7. So, if a word is a seed word it will be assigned the correct sentiment aspect and if it is an opinion word it will be assigned to the correct sentiment aspect. This means that only seed words and opinion words can affect the aspects produced.

3.3.2. Results

Table 8 shows the accuracy results for each of the 3 datasets using various types of prior information. The best results for each type of prior information are highlighted in bold. For all 3 datasets, the best accuracy is achieved using Paradigm++. This indicates that using seed words relating to the dataset achieves the best performance. The reason for this is that domain specific sentiment seed words are used as input to our model. The seed words used are taken from the dataset and are the correct sentiment for this domain, when a seed word in kept in the correct aspect, it encourages other related words to be grouped in that aspect. The subjectivity lexicon also performs well across all datasets, this shows that if manual seed words cannot be obtained, reasonable results can still be achieved. The subjectivity lexicon has considerably more seed words than Paradigm++ but Paradigm++ has higher accuracy, which proves that relevant seed words can have much more impact on performance than the amount of seed words.

On analyzing the effect of incorporating POS tagging, we found that for 2 of the 3 datasets the highest accuracy is achieved when incorporating POS tags into the model. This result provides evidence that taking the context of a sentence into consideration can produce more accurate results. Our modified Gibbs sampling equation for incorporating POS tags may have this improvement on performance as opinion words are highly likely to be only adjectives, verbs and adverbs. Consequently if we only include these types of words when discovering our sentiment aspects, we eliminate the possibility of a word which is not an opinion being included in the discovered aspect. For example, the noun 'dinner' cannot be included in an aspect sentiment as it is neither positive or negative, but the adjectives 'delicious', 'lovely' and 'tasty' can all be included in a positive aspect sentiment.

Table 8. Accuracy of incorporating different prior information.

	Sentiment Prior info	**# of polarity words (pos/neg)**	**TripAdvisor Accuracy**	**Tv Accuracy**	**Mp3 Accuracy**
Enhanced Twofold-LDA	Paradigm	26/20	75.611	70.897	66.047
Enhanced Twofold-LDA	Paradigm ++	119/78	78.653	**72.520**	65.951
Enhanced Twofold-LDA	Subjectivity lexicon	822/1113	75.623	72.240	76.484
Enhanced Twofold-LDA	Paradigm ++ + Pos tags	119/78	**79.261**	70.272	**78.555**
Enhanced Twofold-LDA	Lexicon + POS tags	822/1113	75.628	67.837	77.803

Naïve Bayes	n/a	82.473	86.012	76.705
Lang model	n/a	82.473	86.012	76.705

If we now compare the Enhanced Twofold-LDA model to the state-of-art classification techniques, naïve Bayes and language model, we can see that for the Mp3 dataset, the Enhanced Twofold-LDA model achieves the highest accuracy. For the TripAdvisor dataset, the Enhanced Twofold-LDA model is comparable. Lastly, for the TV dataset, Naive Bayes and language model are noticeably better, possibly due to the fact that the TripAdvisor dataset is the largest and as the two state-of-art methods require training data, there is considerably more training data with this dataset. These results can be considered reasonably good. With the Joint Sentiment/Topic model, none of the experiments reached results as high as state-of-art techniques [17]. Naive Bayes and language model may have better accuracy in some cases but they both require a large amount of training data in order to achieve good results. The Enhanced Twofold-LDA model requires no training data. It does require prior information, however results show that for sentiment, general seed words which are not manually obtained can still achieve respectable performance. Another point to add is that Naive Bayes and language model take considerably longer to process, especially with a large dataset. The language model took up to 2 days to process results for the TripAdvisor dataset, with Twofold-LDA model it only took a couple of hours.

Naive Bayes and language model are well known for achieving good performance for the task of sentiment classification. As we were unable to produce results as high as these state-of-art techniques, we look at their learning processes in more detail to understand possible reasons. Naive Bayes is a simple probabilistic classifier which often outperforms more sophisticated models [22]. First, we have the prior information of the training data which is the overall probability of positive and negative classes, and then have the posterior probability that each word appears in a class. Finally, we multiply the positive prior and the probability that all the words are positive and compare this against the negative prior and the probability that all the words are negative, this sentence will be classified as the class with the highest probability.

A language model is simply the probability distribution of generating each word. If we train a language model for positive and a language model for negative, to determine which class a test sentence belongs to, we simply calculate the probability that all words comes from the positive language model, and also calculate the probability that all words are from the negative language model. Again, the sentence will be classified as the class with the highest probability.

The Enhanced Twofold-LDA learning process is considerably different to Naive Bayes and language model which could explain the difference in results for sentiment classification. To train an LDA model we first randomly assign each word in a document with a topic e.g., positive or negative. If we use an LDA model for aspect discovery, there would usually be much more topics than this. This will give a random topic distribution over documents and word distributions over topics. To improve these random distributions, the model goes through each word in a document and assumes that all topic assignments are correct except the one in question, then the amount of words in the sentence currently assigned to the topic e.g., p(pos|sentence) is calculated and the probability that the topic contains the word e.g., p(great|pos), this is essentially the probability that the topic generated the word. The current word is then re-sampled with this probability. This step is repeated a considerable number of times. When sampling is complete, more accurate topic distributions are produced– which is the amount of words assigned to a topic in the document, and word distributions – which is the amount of words assigned to a topic

Naive Bayes and language model are commonly used and achieve high performance in identifying sentiment, which is either positive, negative and sometimes neutral. The LDA model on the other hand was designed to identify numerous topics, often 30-100, therefore, using an LDA model to discover only 2 or 3 topics (positive, negative and neutral) may have an effect on the performance. Both Naive Bayes and language model simply use the probability of words occurring in a specific class in order determine sentiment which proves better than the LDA models method of trying to find topic distributions in a document. Thus, the empirical results reveal that calculating the probability that documents are positive or negative using the probability of words in training documents provides much better performance than calculating the probability that a document is positive or negative using re-sampled topic distributions in a set of documents.

3.3.3. Further investigation of state-of-the-art techniques

An interesting observation is that the results for Naive Bayes and language model are the same for each dataset as shown in Table 8. We decided to further investigate this and found that this is a result of converting the sentence-level results to review-level. For both Naive Bayes and language model, there are more sentences classified positive than negative for every review. Therefore, at review-level all reviews are classified as positive. Our results show that for Naive Bayes and language model, 100% of the reviews were classified as positive and 82.473% of the reviews are positive, therefore both methods have an accuracy of 82.473% as shown in Table 8. With the Enhanced Twofold-LDA model, results show that at sentence-level a review can be classified with more negative than positive sentences meaning that at review-level we see a mixture of positive and negative reviews.

We therefore decided to carry out some additional experiments, taking the positive and negative intermediate outputs from the Enhanced Twofold-LDA model and using them as training data for Naive Bayes and language model. This will allow for the comparison of manually labelled data and data labelled by the topic model. A reasonable assumption would be that the manually labelled data will perform better than the data labelled by the topic model.

Table 9 shows the accuracy results of Naive Bayes (NB) and language model (LM) using the results output by the Enhanced Twofold-LDA model. Again we see that the accuracy is the same across the datasets for both Naive Bayes and language model due to the high number of classified positive sentences in each review i.e., in every review there are more positive sentences than negative. The results in Table 9 show that for the Mp3 dataset using the intermediate outputs from the Enhanced Twofold-LDA model has higher accuracy than the original dataset. This provides additional evidence that the Enhanced Twofold-LDA model can provide better performance than state-of-art techniques. For TripAdvisor and TV datasets however, using the intermediate results decreases the performance. Therefore this experiment shows the majority of datasets show a higher accuracy with manually labelled data in comparison to data labelled by the topic model.

If we compare these datasets, one difference in the Mp3 dataset in comparison to the other two is that on average this dataset contains much shorter sentences which may be a contributing factor to these results, for instance, a short sentence may achieve higher sentiment classification accuracy than longer sentences. Both TV and Mp3 would have similar content in describing a product. TripAdvisor would contain details such as the weather or location. Since TV and Mp3 have similar content but opposite affects on the results we can presume that the content i.e., opinions on electronic products, does not affect sentiment classification. An example of content affecting sentiment classification would be movie reviews, a positive review about a horror movie could be classified as negative due to content on scary or violent scenes.

Table 9. NB and LM using Enhanced Twofold-LDA results

90/10	**TripAdvisor**	**Tv**	**Mp3**
Naïve Bayes	**82.473**	**86.012**	76.705
NB using Enhanced Twofold-LDA output	81.803	78.862	**78.555**
Langauge model	**82.473**	**86.012**	76.705
LM using Enhanced Twofold-LDA output	81.803	78.862	**78.555**

3.3.4. Comparison of the Enhanced Twofold-LDA model and ASUM.

Next, we make a comparison on the performance of the Enhanced Twofold-LDA model and ASUM to investigate sentiment classification. Again the three datasets were used for experiments. Each model discovers 2 classes of sentiment, one positive and one negative. To calculate the baseline, the PARADIGM++ seed words are used. Each sentence is labeled according to the highest number of positive or negative seed words it contains. For experiments, the two models first classify each sentence as either positive or negative. The sentence-level results are then aggregated to review-level results by taking the highest probability of sentences, for example, reviews with a higher number of negative sentences are classified as negative, and reviews with a higher number of positive sentences or equal to the number of negative sentences, are classified as positive.

We use the best results for the Enhanced Twofold-LDA model taken from Table 8, i.e., Paradigm++ and POS tags for the TripAdvisor dataset, Paradigm++ for the TV (aspect sentences) dataset and Paradigm++ and POS tags for the Mp3 dataset. ASUM uses Paradigm++ as prior knowledge via sentiment seed words.

Table 10: Comparison of sentiment accuracy for the Enhanced Twofold-LDA model and ASUM.

	TripAdvisor Accuracy	**Tv Accuracy**	**Mp3 Accuracy**
Enhanced Twofold-LDA	79.261	72.520	**78.555**
ASUM	**82.473**	**86.05**	74.826
Baseline	**82.473**	55.028	76.705

Table 10 shows the sentiment classification accuracy of the two models and the baseline. The highest accuracy for each of the three datasets is highlighted in bold. The Enhanced Twofold-LDA model has the highest accuracy for the Mp3 dataset, whereas ASUM shows a higher accuracy for the TV and TripAdvisor datasets. The accuracy of the Enhanced Twofold-LDA model is slightly lower for the TripAdvisor dataset and significantly lower than ASUM for the TV dataset. An interesting observation is that the baseline also performs poorly for the TV dataset. This would indicate that the seed words in the reviews do not match the sentiment of the review. Another possible reason would be that the sentiment seed words may not be frequently used throughout the sentences. The Enhanced Twofold-LDA model achieves improved accuracy with a high frequency of seed words, therefore the possible low frequency of seed words may be a contributing factor to the poor performance for the TV dataset. As with naïve Bayes and language model in the previous section, the ASUM results for TripAdvisor and TV show the same pattern as all sentences are classified as positive at review level, therefore the accuracy is equivalent to the percentage of total positive reviews.

ASUM outperforms the Twofold-LDA model in two out of the three, therefore an analysis is made on the differences of the two models as both are based on the LDA model. Also possible factors are discussed for the reasoning behind ASUM achieves higher sentiment classification accuracy.

- Difference between models
  The two models have one main difference; ASUM jointly models aspects and sentiment, whereas the Twofold-LDA model discovers aspect and sentiment separately. ASUM models the sentiments towards different aspects, this results in aspects which are closely related to sentiments. Table 10 indicates that jointly modeling aspects and sentiments together achieves higher sentiment classification accuracy.

- Sentiment classification methods
  ASUM incorporates prior information, in the form of sentiment seed words, into the LDA model using asymmetric β. The Twofold-LDA model on the

other hand incorporates sentiment seed words into the model by altering topic labels z. This indicates that using prior probabilities may be a better approach to exploiting sentiment in comparison to altering topic labels.

- Use of seed words
  The use of seed words is not a contributing factor for the difference in performance as the same set of sentiment seed words are used for both ASUM and the Enhanced Twofold-LDA model.

- Possible over-fitting
  The results in section 3.2.1 show that ASUM is unable to discover negative seed words within the negative *Room* topic. This is reflected in Table 7 as all reviews are classified as positive.

As a result of this investigation, a future study on the Enhanced Twofold-LDA model will include improving how the model does sentiment classification. From the points above, the first place to start will be jointly modeling aspect and sentiment together so as to produce aspects which directly relate to sentiments. Another direction will be to study Dirichlet priors and find various ways to alter the distributions produced by the model.

3.3.5. Experiment with all sentences

Removing sentences with no seed words reduces the coverage of the dataset, therefore an additional experiment is performed on the TV (all sentences) dataset to ensure removing these sentences do not affect sentiment classification results. Table 11 shows a comparison between the classification accuracy of the TV (aspect sentences) and TV (all sentences) datasets. Accuracy is increased for a number of experiments with Paradgim++ showing the biggest increase. Subjectivity lexicon and Paradigm++ with POS tags show a slight decrease. This experiment proves that the previous experiments which use only sentences with aspect seed words can effectively determine the sentiment expressed for those reviews. These subsets of sentences are representative of the overall sentiment.

Table 11: Comparison of sentiment accuracy for TV (aspect sentences) and TV (all sentences) dataset.

Sentiment Prior info	# of polarity words (pos/neg)	TY (aspect sentences) Accuracy	TV (all sentences) Accuracy
Paradigm	26/20	70.897	75.72
Paradigm ++	119/78	**72.520**	**80.66**
Subjectivity lexicon (only strong)	822/1113	72.240	71.39
Paradigm ++ & POS tags	119/78	70.272	69.06
Lexicon & POS tags	822/1113	67.837	67.846

3.4. *The Enhanced Twofold-LDA Model*

One aim of this paper was to improve performance and efficiency of the Twofold-LDA model. To do this we proposed the Enhanced Twofold-LDA model which has the ability to automatically provide the same graphical output as the Twofold-LDA model, which takes much less time and effort in comparison to manually creating the graph. To evaluate the computational time and effort it takes to create the chart manually using the Twofold-LDA model and automatically using the Enhanced Twofold-LDA model, we calculated the time it took for creating a chart for the TripAdvisor, TV and Mp3 datasets.

Table 12. Time taken to complete each task in minutes

Dataset	Model	Enhanced Twofold	Twofold aspect	Twofold sentiment	Produce graph (approx)	Total time
Mp3	Enhanced Twofold-LDA	6.57	-	-	-	6.57
	Twofold-LDA	-	3.09	3.46	9.00	15.55
TV	Enhanced Twofold-LDA	6.21	-	-	-	6.21
	Twofold-LDA	-	3.01	3.33	9.00	15.34
TripAdvisor	Enhanced Twofold-LDA	127.24	-	-	-	127.24
	Twofold-LDA	-	68.57	66.48	9.00	144.05

The results for the 3 datasets are shown in Table 12. We can see that for the Mp3 and TV datasets, it takes over 1.5 more time to produce the graph manually using the

Twofold-LDA model. As the TripAdvisor dataset takes considerably more time to run, producing the graph does not have as big impact on the overall time nevertheless, the Enhanced Twofold-LDA model is quicker at producing the graph. These results prove that our goal of making the Enhanced Twofold-LDA model more efficient has been succeeded.

Figure 9 shows an example of the automatic output produced by the TripAdvisor dataset. Here we illustrate that the Enhanced Twofold-LDA model can produce the same graphical chart as the Twofold-LDA model with a lot less effort and time. A customer or manufacturer can see clearly the opinions expressed towards each aspect. *Checkin* is the most talked about aspect with mostly positive views. *Location* has received no negative comments and *Value* has an equal number of positive and negative comments although not many people comment on *Value* so there is not as high a confidence in this opinion than there would be with the views on *Room*, here there is also about an even number of positive and negative comments but as more people have commented this makes the confidence in these opinions higher.

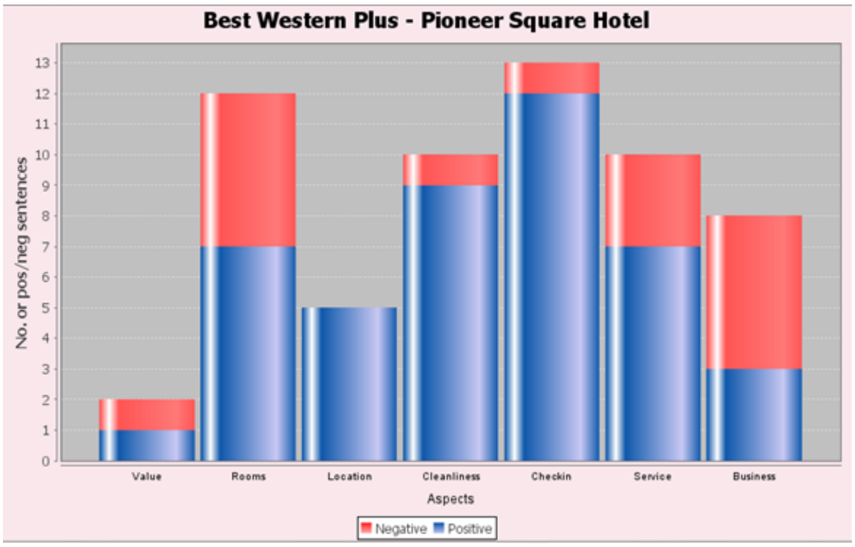


Fig. 9. Example graphical output of hotel review

3.4.1. *Investigating Charts*

The Enhanced Twofold-LDA model automatically outputs stacked vertical bar charts of results, indicating the positive and negative opinion towards each aspect. These charts are at benefiting to end users, therefore a further investigation is carried out on customer satisfaction towards different types of charts. We give out questionnaire to a sample of 23 people to compare 3 Mp3 players using 5 different types of charts. The questionnaire was designed to obtain the following information:

- Best and worst Mp3 for each chart.
- How easy the chart was for comparing Mp3 players.
- Which chart type was the easiest for comparison?
- Which chart type was the most difficult for comparison?

Figures 10(a)-10(e) show an example of each type of chart; Standard vertical bar chart, Stacked vertical bar chart, Standard horizontal bar chart, Stacked horizontal bar chart and In line vertical bar chart. On a scale of 1 to 10, the sample of end users were asked to score each type of chart on how easy the chart is for comparing aspects. Also, the sample were asked to score the ease of use for the charts, this was how easy it was to make a decision on which Mp3 was the best and worst. The results on the ease of use scores are shown on a bar chart to the right of Figures 10a-10e.

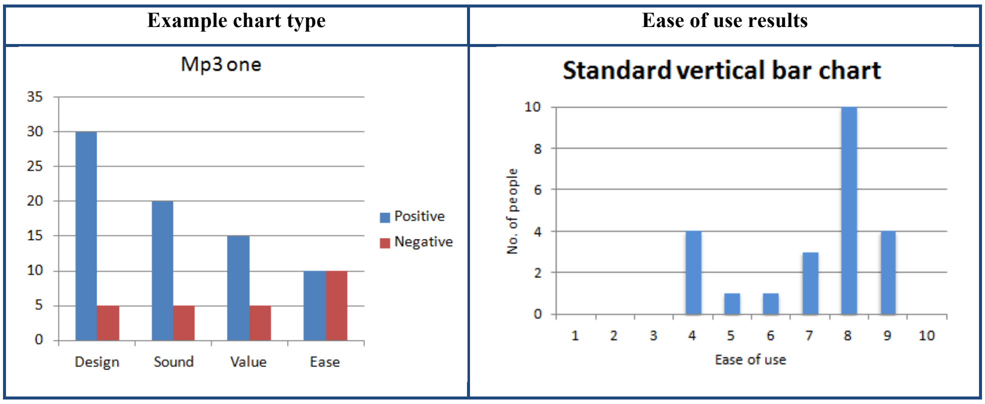


Fig. 10(a): Standard vertical bar chart.

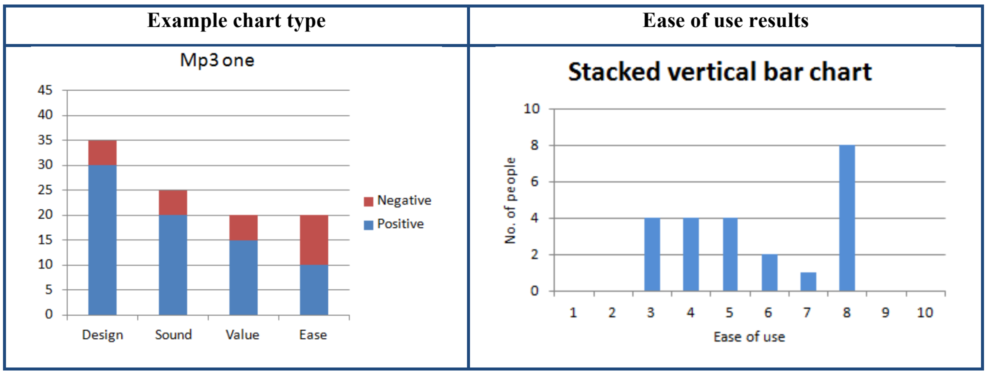


Fig. 10(b): Stacked vertical bar chart.

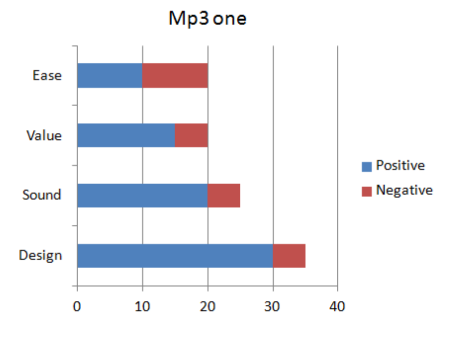
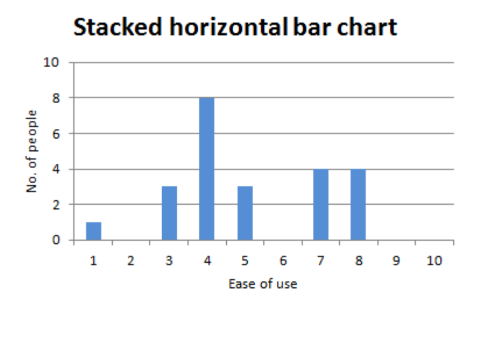
Example chart type	Ease of use results
	

Fig. 10(c): Standard horizontal bar chart.

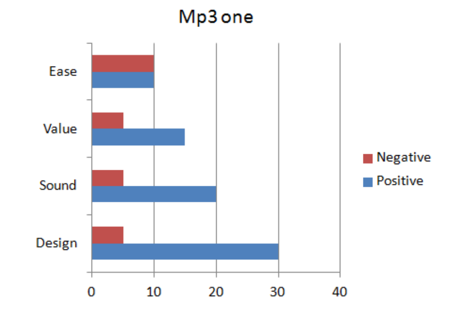
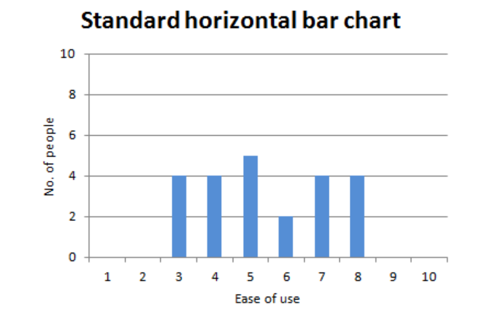
Example chart type	Ease of use results
	

Fig. 10(d): Stacked horizontal bar chart.

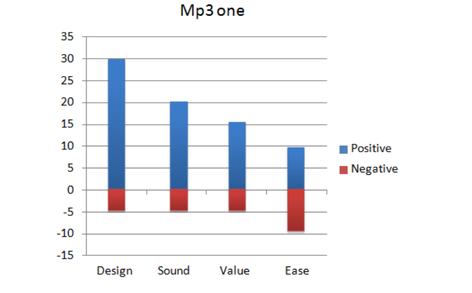
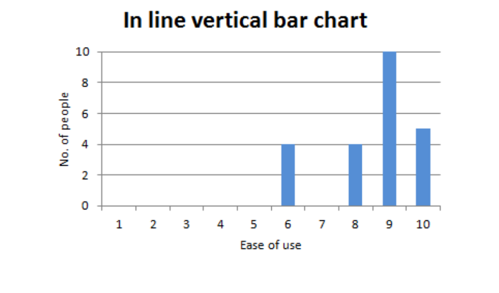
Example chart type	Ease of use results
	

Fig. 10(e): In line vertical bar chart.

The findings from the questionnaire indicate the stacked vertical bar chart used for the Enhanced Twofold-LDA model in Figure 9 is considered about average in terms of how easy the chart is to use for comparing aspects. The findings in this study indicate the importance of research for customer satisfaction as people may have different views on different graphical outputs. For our Enhanced Twofold-LDA model, it is important that end users find the chart easy to compare products and make a decision on which product is the best. Figures 10(a)-10(e) reveal the following observations for ease of use:

- Standard vertical bar chart – generally good, above average.
- Stacked vertical bar chart – okay, about average.
- Standard horizontal bar chart – okay, about average.
- Stacked horizontal bar chart – mixed results, generally poor.
- In line vertical bar chart – all very good, best overall.

Figure 11 shows the results for the chart that the sample liked best and least for comparing aspects and overall products. These results reflect the observations which are made in the points above. Figure 11 clearly shows that the In line vertical bar chart is the preferred chart for all 23 end users in the sample.

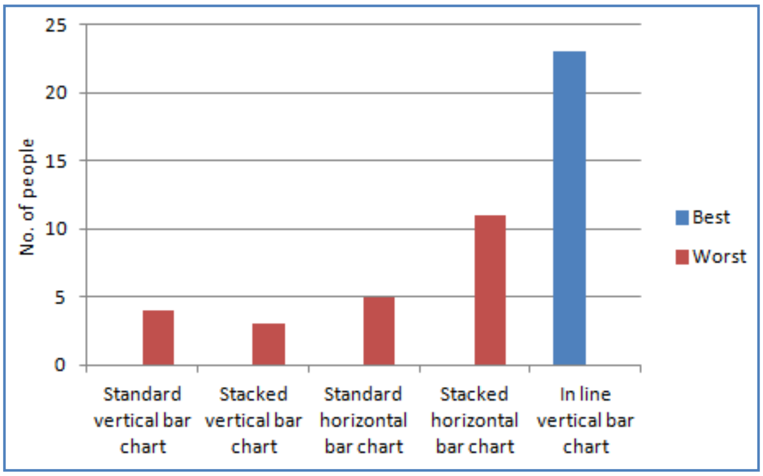


Fig. 11: Best and worst charts.

Finally, an interesting finding is made from analyzing which Mp3 is considered best and worst by the sample. *Mp3 one* and *Mp3 two* have very similar results:

- *Mp3 one* has 75 positive reviews and 35 negative reviews.
- *Mp3 two* has 70 positive reviews and 30 negative review.

When creating the questionnaire the three Mp3 players are placed in different order for each test. This prevented the sample from predicting the outcome of the next test. The majority of the sample did not select the same two best and worst Mp3 players for each test. For example, one user has mixed results for *Mp3 one* and *Mp3 two*, the similar content prevents the user from making the same decision in each test. In some cases the user found *Mp3 one* worst and in some cases *Mp3 two* worst. This is another indication that the right chart which shows results clearly is a very important factor to the decision of end users.

4. Related Work

Our work focuses on three topics: aspect discovery, sentiment analysis and jointly modeling aspect and sentiment. We therefore review previous research on these topics.

4.1. *Aspect discovery*

Aspect discovery is the process of finding aspects relating to a product or service in a piece of text. It could be finding the main topics in a new article or finding the most talked parts in a set of movie reviews. It is a very powerful tool to make sense of large unlabelled datasets. One popular method for aspect discovery is to use association rules. Association rule mining is used in [14, 15], they use POS tagging to extract noun phrases which are used as candidate aspects, firstly they find all the frequently occurring aspects by using frequent itemsets as candidate aspects then pruning the results, removing uninteresting or redundant aspects. They also investigate infrequent aspects, which are not repeated throughout but still could be useful. They take sentences containing opinion words but no frequent aspects, use the closest noun/noun phase and then add them to the aspect set. Other work includes various rules to find information about the product, its aspects and opinions [21]. The different rules can identify the opinions towards products, the aspects of the product and the opinions towards the aspects. A number of sentiment analysis systems have been developed for discovery aspects, these include Opinion observer [5], OPINE [23], and aggregating aspects by collaborative topic regression [44, 45]. More recently the relationship between aspects and opinions have been studied in [35, 42, 43], because opinions are generally expressed on some aspects in online reviews. Such dependency relationship was first used to find the nearest nouns and noun phrases of a set of known opinion words.

Another approach for aspect discovery is topic-modeling [24]. It has become common to use the Latent Dirichlet Allocation model [2] to discover latent aspects in a set of unlabeled data, which is similar to the semi-supervised approach [46]. There have been many variations and extensions to the LDA model. For instance, prior information has been incorporated into the model to add supervision [25, 26, 27]. LDA has also been extended in many forms, for instance by adding a subjectivity layer [28], adding multi-gain topics [29], restricting one-to-one correspondence between aspects and user tags [30], using tag-topic models for mining blog data [31] and finally providing review summarization [33].

Our work is similar to some previous research, which they guide our output with prior knowledge, we however use seed words for both aspect and sentiment to produce relevant results which can be transformed to visualize the overall results in a user-friendly chart. The previous research has been carried out to try and provide a more useful LDA output, they however require aspect labels, and the output they provide is a set of structured sentences [32]. Our output is a chart, which clearly shows the opinion towards each aspect that can be interrupted with minimum effort, as opposed to a set of structured sentences that would still require effort to read and interrupt. Another work which has investigated visualizing results is one an application for comparing topic models [33]. The application can compare conceptual content and document relationships, but our work differs in that we want to provide a visual output for customers or manufacturers rather than analyzing the topic model itself. A final paper which probably solves the most similar problem to our work is one as in [8], they generate a rated aspect summary of short comments. The major difference between their work and ours is that they only deal with short comments, therefore they take a different approach to finding aspects and sentiment. In short comments, an opinion expressed on an aspect is usually a concise phrase e.g., 'bad picture' or 'well designed'. We deal with larger free-format text which can be much more difficult to evaluate as comments on aspect and sentiment are expressed in many different ways.

4.2. *Sentiment classification*

Machine learning techniques have been commonly used as a means of evaluating sentiment. Earlier work classifies sentiment at document-level, where a whole document are classified as positive, negative or possibly neutral. Much research has been carried out on the analysis of machine learning techniques, a popular study investigates a number of techniques on movie reviews and concludes that SVMs produce the best results [20]. Previous research calculates the mutual information between a phrase and reference words to indicate the sentiment of the phrase, the calculation rating can also be used to indicate the strength of the semantic orientation [34, 37]. There are several survey publications about this topic, which summarizes the general sentiment analysis problems and methods [36], and deals with cross-lingual sentiment analysis [19]. We show the strength of sentiment orientation by indicating the amount of times that an aspect has been described as positive or negative, the more an aspect has been described as positive for example, the more confidence we can have in this opinion.

Sentiment may also be classified at sentence-level [34, 38], the authors use a similar approach to one in which they measure the similarity between pair of words or phrases with the added assumption that opinions of opposite sentiment are inclined not to appear together at sentence-level. Their experiments at sentence-level prove that performance is improved in comparison to those at document-level. Our work considers sentiment at sentence-level as we wish to identify the aspect(s) and sentiment of each sentence that contains an aspect.

Topic modeling may also be used to discover sentiment. Using topic models for finding sentiment however, is harder than topic-based classification as sentiment is expressed in more subtle ways and can sometimes be domain specific [17, 18, 20]. Incorporating prior information into models can be an effective way for discovering sentiment [4, 17]. Our research therefore investigates topic modeling for sentiment and how seed words can improve the performance of sentiment classification.

An observation was made that natural language processing approaches to sentiment classification achieve high performance and often use POS tagging as a means of finding out the meaning of each word in a sentence [34, 37]. They calculate the average semantic orientation of phrases in a review containing adjectives or adverbs. Additionally, they use adjectives as opinion words and applies association rules to the dataset to find opinions about products. Finally, in studies [14, 15, 21], the authors extract the nearby adjective from a sentence containing an aspect word. These studies demonstrate that using POS tags can help understand the meaning of words in a sentence and can encourage better performance. We therefore wish to incorporate POS tagging from natural language processing techniques into topic modeling so as to improve performance and create an Enhanced Twofold-LDA model. To the best of our knowledge this has never been done before. We will also apply this same technique to our aspect model and analyze the results.

4.3. *Jointly modeling aspect and sentiment*

A number of unified models have been proposed for extracting both aspect and sentiment [12, 13, 39, 40]. Some of the more popular models include:

Latent Aspect Rating Analysis (LARA) [7] defines the problem of analyzing opinions expressed about aspects in online reviews at the level of topical aspects. They propose Latent Rating Regression (LRR) model which uses the overall rating given by a reviewer to discover the latent ratings on each aspect and the weight given to each aspect for the overall judgment. Aspect seed words are used in order to help discover aspects similar to Enhanced Twofold-LDA. LRR requires all reviews to have an overall rating provided, whereas we require no labeled data for our Enhanced Twofold-LDA model. Follow up research was carried out with an extension that requires no aspect seed word supervision however, they still require labeled data [7]. LRR does not include any form of displaying results in their work, while our research provides a chart which is very useful for end-users.

Joint Sentiment/Topic (JST) [17, 18] is a fully unsupervised model which adds a sentiment layer to LDA in order to detect aspect and sentiment simultaneously. JST uses seed words in order to detect sentiment, Enhanced Twofold-LDA differs in that we use seed words to discover both aspect and sentiment. Another difference is that we detect sentiment at sentence-level whereas JST detects sentiment at document-level. A document-level sentiment will be much more generic than sentence-level sentiments. JST also provides no form of displaying results, when analyzing customer reviews it can be seen as very beneficial to be able to show customers the findings of our results.

Aspect and Sentiment Unification Model (ASUM) [4] incorporates sentiment into the unified model so that the resulting model will signify the probability distributions over words for pairs of aspect and sentiment. Both the Enhanced Twofold-LDA model and ASUM use seed words for identifying sentiment. The Enhanced Twofold-LDA model differs in that ASUM outputs senti-aspects containing both aspect and sentiment, whereas we output separate aspect and sentiment so that we can combine the results in visual form which will be shown later in the paper. We also incorporate POS tags to help discover sentiment. ASUM provides the output of results for restaurant reviews as shown in Figure 3. An end-user can see the different aspects and the opinion expressed towards each aspect but the figure is not user friendly and would not be easy to use to compare with other restaurants.

5. Conclusion

In this paper we have provided an analysis into the Twofold-LDA model and developed the Enhanced Twofold-LDA model which incorporates natural language processing techniques into this model that can automatically determine aspect and sentiment in graphical form, whereby creating a much more efficient method. An additional investigation has been also carried out on various types of graphical output that reveal which charts end users preferred. For aspect discovery, the experiments demonstrate that the Enhanced Twofold-LDA model is able to produce aspects more closely related to aspects than those produced by ASUM, while achieving the highest performance for 1 of the 3 datasets in comparison to ASUM. We then investigated sentiment classification and different prior information and found that seed words relevant to the dataset are more effective than general seed words. Finally, we looked at comparing the Twofold-LDA model against the Enhanced Twofold-LDA model and found that for 2 or the 3 datasets, it took 1.5 more time to produce the graph with the Twofold-LDA model as the graph is created manually after the results are obtained. This clarifies that the Enhanced Twofold-LDA model is much more efficient. An additional investigation also shows that the graph produced by the Enhanced Twofold-LDA model, Stacked vertical bar chart, was rated average for ease of use by a sample of end users. The questionnaire analysis reveals that a vertical bar chart is preferred for end users as it is easy for comparing different aspects.

In conclusion, we have achieved each of the aims we set out to do. Firstly, we improved the efficiency of the Twofold-LDA model by automatically producing a chart. Next, we improved the performance of sentiment classification by incorporating part-of-speech tagging into the sampling process. Finally, we compared the proposed Enhanced Twofold-LDA model with a recent comparable study, namely ASUM.

References

1. Burns, N., Bi, Y., Wang, H. & Anderson, T. A Twofold-LDA Model for Customer Review Analysis, In: *Proceedings of Web Intelligence and Intelligent Agent Technology. (*ACM, 2011), pp. 253 - 256.
2. Blei, D.M., Ng, Y.A. & Jordan, M.I., Latent Dirichlet Allocation. *Journal of Machine Learning Research,* 3(2003), 993-1022.
3. Liu, B., Hu, M. & Cheng, J., Opinion Observer: Analyzing and Comparing Opinions on the Web, In: *Proceedings of the 14th international conference on World Wide Web. (*ACM, New York, NY, USA, 2005), pp. 342-351.
4. Jo, Y. & Oh, A., Aspect and Sentiment Unification Model for Online Review Analysis", In*: Proceedings of the fourth ACM international conference on Web search and data mining. (*ACM, New York, 2011).
5. Andrzejewski, D. & Zhu, X., Latent Dirichlet Allocation with topic-in-set knowledge", In: *NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing. (*Association for Computational Linguistics, Stroudsburg, PA, USA, 2009) pp. 43–48.
6. Griffiths, T.L. & Steyvers, M., Finding Scientific Topics. In: *Proceedings of the National Academy of Sciences of the United States of America*, 101 (2004), 5228-5235.
7. Wang, H., Lu, Y. & Zhai, C., Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach, In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. (*ACM, New York, NY, USA, 2010) pp. 783-792.
8. Lu, Y., Zhai, C. & Sundaresan, N., Rated Aspect Summarization of Short Comments, In*: Proceedings of the 18th international conference on World wide web. (*ACM, New York, NY, USA, 2009) pp. 131-140.
9. Burns, N., Bi, Y., Wang, H. & Anderson, T., Extended Twofold-LDA Model for Two Aspects in One Sentence. In: *14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. (Springer Berlin Heidelberg, 2012) pp 265-275.
10. Yan, X., Guo, J., Lan, Y. and Cheng, X. A Biterm Topic Model for Short Texts. In Proceedings of the 22nd International Conference on World Wide Web, pages 1445–1456, 2013.
11. Brody, S. & Elhadad, N., An Unsupervised Aspect-Sentiment Model for Online Reviews, In: *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics.* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2010) pp. 804-812.
12. Zhao, W.X., Jiang, J. & Yan, H.L., X., Jointly Modeling Aspects and Opinions with a MaxEnt-LDA Hybrid", In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. (*Association for Computational Linguistics, Stroudsburg, PA, USA, 2010) pp. 56-65.
13. Zhu et al 2011, Aspect-Based Opinion Polling from Customer Reviews, In: *IEEE Transactions on Affective Computing*. Vol 2 (IEEE, 2011), pp 37-49.
14. Hu, M. & Liu, B., Mining and summarizing customer reviews, In: *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining. (*ACM, New York, NY, USA, 2004a) pp. 168 - 177.
15. Hu, M. & Liu, B., Mining opinion features in customer reviews, In: *19th national conference on Artificial intelligence. (*AAAI Press / The MIT Press, 2004b) pp. 755-760.
16. Chen, B., Zhu, L., Kifer, D. & Lee, D., What is an Opinion About? Exploring Political Standpoints Using Opinion Scoring Model, In: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence. (*AAAI press, 2010).
17. Lin, C. & He, Y., Joint Sentiment/Topic Model for Sentiment Analysis, In: *Proceedings of the Conference on Information and Knowledge Management. (*ACM, New York, 2009), pp. 375-384.

18. Lin, C., He, Y. & Everson, R., Sentence Subjectivity Detection with Weakly-Supervised Learning, In: *The 5th International Joint Conference on Natural Language Processing*, (2011) pp. 1153–1161.
19. Feldman R (2013) Techniques and applications for sentiment analysis. Communications of the ACM 56(4):82, DOI 10.1145/2436256.2436274
20. Pang, B., Lee, L. & Vaithyanathan, S., Thumbs up? Sentiment Classification using Machine Learning Techniques, In: *ACL-02 conference on Empirical methods in natural language processing - Volume 10. (*Association for Computational Linguistics, Morristown, NJ, USA, 2002) pp. 79 - 86.
21. Kim, W., Ryu, J., Kim, K. & Kim, U., A Method for Opinion Mining of Product Reviews using Association Rules, In: *Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human.* ***(***ACM, New York, NY, USA, 2009) pp. 270-274.
22. StatSoft, I. 2012*, Naive Bayes Classifier*. Available: http://www.statsoft.com/textbook/Naive-bayes-classifier/ [28/11/2012] .
23. Popescu, A. & Etzioni, O., Extracting Product Features and Opinions from Reviews, In: *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. (*Association for Computational Linguistics, Morristown, NJ, USA, 2005) pp. 339 - 346.
24. Blei, D.M. & McAuliffe, J.D., Supervised Topic Models, In: *Advances in Neural Information Processing Systems*. (2007)
25. Andrzejewski, D., Zhu, X. & Craven, M., Incorporating Domain Knowledge into Topic Modelling via Dirichlet Forest Priors, In: *Proceedings of the 26th Annual International Conference on Machine Learning. (*ACM, New York, 2009).
26. Andrzejewski, D., Zhu, X., Carven, M. & Recht, B., A Framework for Incorporating General Domain Knowledge into Latent Dirichlet Allocation using. First-Order Logic, In: *Proceedings of the 22nd International Joint Conference on Artificial Intelligence* (2011).
27. Lakshminarayanan, B. & Raich, R., Inference in Supervised Latent Dirichlet Allocation, *In International Workshop on Machine Learning for Signal Processing. (*IEEE, 2011), pp. 1 - 6.
28. Lin, C., He, Y., Everson, R. & Ruger, S., Weakly Supervised Joint Sentiment-Topic Detection from Text, In: *IEEE Transactions on Knowledge and Data Engineering,* (2011) pp. 1134-1145.
29. Titov, I. & McDonald, R., Modeling Online Reviews with Multi-Grain Topic Models, In: *Proceeding of the 17th international conference on World Wide Web. (*ACM, New York, 2008) pp. 111-120.
30. Ramage, D., Hall, D., Nallapati, R. & Manning, C.D., Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora, In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing.(*Association for Computational Linguistics, Stroudsburg, PA, USA, 2009) pp. 248–256.
31. Tsai, F., A tag-topic model for blog mining." *Expert Systems with Applications: An International Journal,* 38(5), (2011) pp.5330–5335.
32. Jin, F., Huang, M. & Zhu, X., Guided Structure-Aware Review Summarization. *Journal of Computer Science and Technology,* 26(4), (2011) 676-684.
33. Crossno, P.J., Wilson, T.M. & Dunlavy, D.M., TopicView: Visually Comparing Topic Models of Text Collections, In: *23rd Conference on Tools with Artificial Intelligence. (*IEEE, 2011) pp. 936-943.
34. Turney, P.D., Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, In: *Proceedings of the 40th Annual Meeting of the ACL. (*Association for Computational Linguistics, Morristown, NJ, USA, 2002) pp. 417 - 424.

35. Poria S, Cambria E, Ku LW, Gui C, Gelbukh A (2014) A Rule-Based Approach to Aspect Extraction from Product Reviews. In: Proceedings of the Second Work- shop on Natural Language Processing for Social Media (SocialNLP), Association for Computational Linguistics and Dublin City University, Dublin, pp 28–37.
36. Ravi K, Ravi V (2015) A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems 89:14–46, DOI 10.1016/ j.knosys.2015.06.015
37. Turney, P.D., Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL, In: *12th European Conference on Machine Learning. (*Springer-Verlag, London, UK, 2001) pp. 491 - 502.
38. Gamon, M. & Aue, A., Automatic identification of sentiment vocabulary exploiting low association with known sentiment terms, In: *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing. (*Association for Computational Linguistics, Morristown, NJ, USA, 2005) pp. 57-64.
39. Mei, Q., Ling, X., Wondra, M., Su, H. & Zhai, C., Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs, In: *Proceedings of the 16th Internationl Conference on World Wide Web. (*ACM, New York, NY, USA, 2007) pp. 171 - 180.
40. Titov, I. & McDonald, R., A Joint Model of Text and Aspect Ratings for Sentiment Summarization", In: *Proceedings of ACL-08* , (2008) pp. 308-316.
41. Wang, H., Lu, Y. & Zhai, C., Latent aspect rating analysis without aspect keyword supervision, In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. (*ACM, New York, NY, USA, 2011) pp. 618-626.
42. Li S, Zhou L, Li Y (2015) Improving aspect extraction by augmenting a frequency- based method with web-based similarity measures. Information Processing & Management 51(1):58–67, DOI 10.1016/j.ipm.2014.08.005
43. Wang, Y. and Guo, Q. Multi-LDA Hybrid Topic Model with Boosting Strategy and its Application in Text Classification. In 33rd Chinese Control Conference, pages 4802–4806, 2014.
44. X. Quan, C. Kit, Y. Ge, and S. J. Pan. Short and Sparse Text Topic Modeling via Self-Aggregation. In Proceedings of the 24th International Joint Conference on Artificial Intelligence, pages 2270–2276, 2015.
45. H. Wu, K. Yue, Y. Pei, B. Li, Y. Zhao, and F. Dong. Collaborative Topic Regression with social trust ensemble for recommendation in social media systems. Knowledge-Based Systems, 97:111–122, 2016.
46. Berna Altınel, Murat Can Ganiz. A new hybrid semi-supervised algorithm for text classification with class-based semantics. Knowledge-Based Systems. 108: 50-64, 2016.
47. Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.
48. Witten, I. H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco, 3rd edition edn.(2011).