Distantly Supervised Aspect Clustering And Naming For E-Commerce Reviews

Prateek Sircar
India Machine Learning
Amazon
sircarp@amazon.com

Aniket Chakrabarti*
Alexa AI
Amazon
chakanik@amazon.com

Deepak Gupta
India Machine Learning
Amazon
dgupt@amazon.com

Anirban Majumdar
India Machine Learning
Amazon
majumda@amazon.com

Abstract

Product aspect extraction from reviews is a critical task for e-commerce services to understand customer preferences and pain points. While aspect phrases extraction and sentiment analysis have received a lot of attention, clustering of aspect phrases and assigning human readable names to clusters in e-commerce reviews is an extremely important and challenging problem due to the scale of the reviews that makes human review infeasible. In this paper, we propose fully automated methods for clustering aspect words and generating human readable names for the clusters without any manually labeled data. We train transformer based sentence embeddings that are aware of unique e-commerce language characteristics (eg. incomplete sentences, spelling and grammar errors, vernacular etc.). We also train transformer based sequence to sequence models to generate human readable aspect names from clusters. Both the models are trained using heuristic based distant supervision. Additionally, the models are used to improve each other. Extensive empirical testing showed that the clustering model improves the Silhouette Score by 64% when compared to the state-of-the-art baseline and the aspect naming model achieves a high ROUGE-L score of 0.79.

1 Introduction

The aspect mining based insights and its polarity extraction from reviews is a critical task for e-commerce services that enables seller to understand fine-grained customer preferences and improve product offerings. Extracting important keywords and analyzing their sentiment is a very well studied area. However, the sheer scale of e-commerce services poses important novel challenges. Firstly, review phrases/keywords about the same aspect category need to be grouped together, since each product may have thousands of reviews and there are millions of products. Such aggregation will enable downstream individual aspect analysis by sellers. Secondly, each review phrases/keyword group needs to be assigned an interpretable aspect name to enable easy analysis. Finally, both steps have to be done without human annotations, as human review at e-commerce scale is infeasible. Note that, in this paper, we would refer to the terms, "phrase", "review phrase" and "snippet" interchangeably to denote subsets of a review text, obtained by splitting a multi-context review into smaller sentences of single context. For example, if review text is "The headphone has a good sound quality but not so good bass quality. It is useful for playing music while working out." then the corresponding review phrases would be "The headphone has a good sound quality", "not so good bass quality" and "It is useful for playing music while working out." We have used some syntactic/lexical rules for context splitting.

For unsupervised aspect grouping, extant methods use clustering (Bancken et al., 2014) (eg. k-means) and topic modeling (Brody and Elhadad, 2010) (eg. LDA) approaches. LDA based topic models assume the words are independently generated given the topic and consequently can't leverage the full context of the review sentences. k-means based techniques can overcome the drawback by using contextual embeddings typically generated by transformer based models (Devlin et al., 2018). However, these general purpose transformer language models fail to capture the nuances of e-commerce reviews' language characteristics, such as code mixed sentences including vernacular, incomplete sentence formation, spelling errors. Consequently, these models fail to generalize to e-commerce domain.Another ma-

*work done while author was at India Machine Learning, Amazon

jor drawback of the LDA/k-means based methods is that these techniques are not able to generate a human interpretable name for the aspects (topics/clusters).

In this paper, we propose a practical framework for grouping aspect phrases from reviews into clusters and generate meaningful aspect names for the clusters at scale without any human labeled data. Specifically, the contributions of this paper are as follows:
(1) The proposed framework is able to cluster reviews into clusters by training a transformer model that is aware of the nuances of e-commerce review language characteristics.
(2) The proposed framework is able to generate human readable aspect names for the clusters by training a transformer based conditional natural language generation model.
(3) The proposed framework uses a heuristic distant supervision, thereby avoiding the need for manually labeled data.

To arrive at aspects, we first cluster the phrases by clustering the phrase embeddings generated by the state-of-the-art general purpose semantic matching SBERT model (Reimers and Gurevych, 2019). We fine-tune the transformer based conditional natural language generation (NLG) model T5 (Raffel et al., 2019) for aspect name generation that is distantly supervised using a heuristic TF-IDF distance based algorithm using the above clustering. Finally, to improve the aspect clustering, we train a transformer on the reviews corpus using masked language model (MLM) and subsequently fine-tune it Siamese style using the pairwise triplet loss. The training data (relevant and irrelevant pairs of phrases) for triplet loss is generated using a novel distant supervision strategy that leverages the earlier clustering output and the name generation model outputs. Consequently, the learned text embeddings are very robust to nuances of the e-commerce reviews domain. We empirically evaluate our framework at scale on reviews from a popular e-commerce service. The distantly supervised semantic embedding based clustering model is able to improve Silhouette Score by 64% over a baseline technique using a state-of-the-art general purpose semantic embedding model. Our distantly supervised aspect name generation model is able to improve the Rouge-L score by 16%.

2 Related Works

Aspect phrase extraction from text corpus is a widely researched topic (Quan and Ren, 2014; Qiu et al., 2011; Zhang et al., 2020; Xu et al., 2019, 2018; Wei et al., 2020; He et al., 2017; Vargas et al., 2020). In this paper we explore two tasks after aspect phrase extraction, (1) aspect grouping into clusters, and (2) aspect name generation, that are specifically important to the e-commerce reviews domain due to its large scale and lack of annotation requiring unsupervised techniques. Aspect grouping is done typically by clustering/topic modeling approaches once the aspect phrases have been extracted. Topic modeling approaches include LDA, pLSA, NMF based aspect extraction (Titov and McDonald, 2008; García-Pablos et al., 2018; Mukherjee and Liu, 2012; Chen et al., 2014; W. Xu and Gong; C. Ding and Peng). A number of clustering approaches have also been explored (Zhai et al., 2010; Chen et al., 2016; Zhai et al., 2011; Bancken et al., 2014; Pessutto et al., 2020). One limitation of extant topic modeling/clustering approaches is that these techniques fail to leverage the semantic context of the entire text while clustering. Recently, pre-trained models capable of capturing contextual representations have been developed (Peters et al., 2018; Devlin et al., 2018). However, vanilla pre-trained embeddings doesn't lead to coherent groupings of aspects as the e-commerce review language is significantly different from general English/web text on which these embeddings models are pre-trained. In this paper, we propose a transformer language embedding model that captures the semantics of e-commerce reviews, thereby leading to robust clustering. Note that our generated embeddings may be used with any existing clustering techniques to improve their quality.

3 Proposed Solution

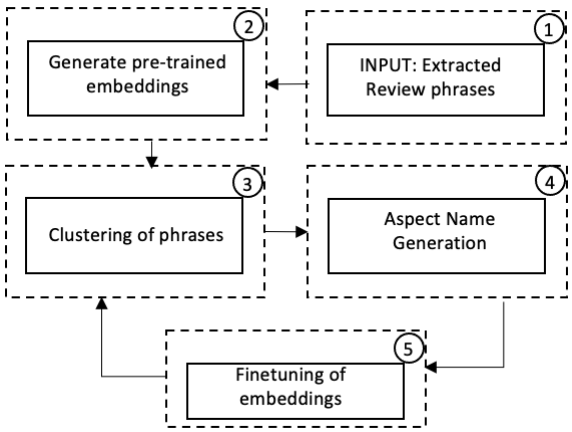
The proposed framework for aspect grouping and naming has two main components: (i) phrase clustering, and (ii) aspect name generation. Our goal is to develop a phrase embedding model that captures the nuances of e-commerce reviews, and a conditional NLG model that generates meaningful names for the aspects without any manually labeled data. To achieve this, we propose a novel distant supervision scheme that uses the two components to improve the other along with some heuristic based automated supervision. Note that

Table 1: Sample output of the aspect insights framework on headphones

Aspect Name	Example Review Snippets
sound quality	['its just like mentioned in description very good quality of sound', 'i must say that i dont regret my decision as its sound quality is too good', 'i can definitely say its sound quality is very good', 'definitely very nice choice its sound is very nice']
value for money	['just go for it on this price bracket it is the complete value for money', 'nothing to dislike as such in this amount of money this is the best thing u get', 'nothing more 1 can ask and to top it all at an amazing price point', 'go for these guys for the price range these are the best']
bass quality	['build quality is pretty good and yeah it does have a punchy bass', 'this must be a nice purchase if you are bass lover', 'just go for it if you are a bass lover', 'just go for it if u are bass lover', 'it is the king of bass so i strongly recommended']

we get the review phrases extracted by the existing pipeline at a popular e-commerce service. Figure 1 shows an overview of the proposed framework

Figure 1: **w**orkflow diagram for clustering and naming



and table 1 shows a snapshot of the final output for a headphone.

3.1 Initial Phrase Clustering

Recently advances in language modeling have resulted in text embedding models (Devlin et al., 2018) such that the embeddings are able to capture the semantics of the text and consequently similar text phrases are mapped to similar vectors. Since our goal is to semantically group review phrases into clusters, we chose the state-of-the-art transformer based semantic text embedding model SBERT-STS (Reimers and Gurevych, 2019) that was trained for the semantic textual similarity (STS) task (Wang et al., 2018). Once each review phrase embedding is generated, we use agglomerative clustering to cluster the review phrases into aspect groups. We chose agglomerative clustering technique instead of k-means as agglomerative clustering is parameterized by only the distance threshold that is easier to tune and interpret in our usecase. While SBERT-STS is a state-of-the-art general purpose semantic embedding model, it fails to generalize to e-commerce reviews. The underlying reason is the nuances of e-commerce reviews, such as the phrases often being short incomplete sentences, presence of code mixed phrases including regional words, presence of spelling and grammar errors. To improve the text embeddings to capture the characteristics of reviews, we propose a novel distant supervision strategy to finetune the SBERT-STS model. We describe this strategy in section 3.3.

3.2 Initial Aspect Name Generation

The goal of this component is to generate a name that represents the common theme of a cluster. We use a sequence-to-sequence based NLG model to generate meaningful aspect names. The main challenge with sequence to sequence models is that they require a significant amount of training data for a stable model. We designed a heuristic based distant supervision strategy that enables us to generate labeled data at scale without human annotation. We chose T5 (Raffel et al., 2019) as the base model as it has been pre-trained on a huge amount of data on multiple NLP tasks, making it a great candidate for transfer learning and stable NLG capabilities. We use k randomly selected review phrases concatenated as the input to T5. We choose the most descriptive n-gram from

a cluster of review phrases that satisfy certain linguistic rules as the distantly supervised label (aspect name) for that cluster as follows:We first collect all n-grams (n=1,2,3,4) from the corpus of reviews in a cluster. Next, we eliminate "ineligible phrases" based on POS-tag based rules. We use SPACY (Honnibal et al., 2020) for POS-tagging. Based on the ngrams, we employ the below rules to eliminate ineligible ngrams. Let t be a ngram whose eligibility we would evaluate. Let pos_t be a set of POS tags for each corresponding word in t. t is an ineligible n-gram if either of the following is satisfied:

1. $len(pos_t) > 1$ and last element of pos_t $\in$ ['DET', 'ADP', 'CCONJ', 'ADV', 'PRON', 'AUX', 'SCONJ', 'PART']
2. $len(pos_t) > 1$ and first element of pos_t $\in$ ['ADV', 'AUX', 'PART', 'PRON', 'ADP', 'CCONJ', 'DET']
3. $pos_t \in$ {['ADP', 'NOUN'], ['ADP', 'PROPN'], ['DET', 'NOUN'],['AUX'], ['ADV'], ['INTJ'], ['DET'], ['VERB'], ['CCONJ']}
4. if first or last word of t is "i".

t is an eligible n-gram overriding the above criteria if either of the following is satisfied:

1. $len(pos_t) > 1$ and last element of $pos_t \in$ ['ADJ'] and first element of $pos_t \in$ ['NOUN','PROPN']
2. First word of t is "not".

For an eligible set of n-grams, we propose the following two heuristic algorithms for training label (cluster name) generation:

(1) TF-IDF-based Naming: We derive TF-IDF scores for each n-gram and weight the TF-IDF score by n (in n-gram), i.e. providing higher weight to longer n-grams. This allows us to get more descriptive names. The candidate n-gram with the highest weighted TF-IDF score is the cluster name.

(2) Distance-Based Naming: For each n-gram we compute the mean cosine distance with each member phrase of the cluster. The n-gram with the minimum distance is considered as the cluster name.

For generating the distantly supervised training labels for our model, we choose high confidence cluster names by setting high thresholds for the aforementioned scores.

3.3 Improving Clustering & Name Generation

Next, we use the initial versions of the reviews phrase clustering and aspect name generation model to distantly supervise and improve each other. One of the main limitations of the initial clustering model was the usage of general purpose semantic embeddings from SBERT-STS that fails to capture the distinct characteristics of e-commerce reviews language. Consequently, many phrases could not be assigned a cluster even though they were relevant to certain aspect of a product and in many cases different clusters were formed for the same aspect. To overcome this limitation, we finetune the transformer based text embedding model with reviews text. We use the unsupervised masked language model (MLM) on the reviews text and couple it with distant supervision signal generated from the T5 based aspect name generation model. Below is the algorithm for training our transformer based text representation model.

We first train the transformer using the standard MLM loss (as described in BERT (Devlin et al., 2018)) on reviews text. This enables the model to learn a robust language model specific to the reviews domain. Furthermore, to enhance the semantic matching capabilities, we finetune our model Siamese style using the following triplet loss:

$$loss = max(||e_a - e_p|| - ||e_a - e_n|| + m, 0) \quad (1)$$

where e_a, e_p and e_n are embeddings of anchor phrase, positive phrase and negative phrase, respectively. m is margin. Negative samples should be at least margin further apart from the anchor than the positive. The anchor and positive phrases refer to the same aspect, whereas anchor and negative phrase refer to different aspects. Minimizing this loss would ensure that embeddings of the phrases mentioned in "anchor phrase" and "positive phrase" are close, while the phrases mentioned in "anchor phrase" and "negative phrase" is far away. The methodology to generate triplet data is described below:

(1) Positive Pairs: We hypothesize that clusters with the same/similar names are talking about the same aspect. Therefore, any randomly selected phrase from one cluster could act as a positive pair for another randomly selected phrase from another. For this, we find the cluster names for each

cluster by leveraging the T5 based aspect name generation model. We also find the medoid of each cluster. Medoid is defined as an element in a cluster which has the least average distance from the remaining elements in the cluster. We use the initial SBERT-STS embeddings to generate embeddings of the cluster names and medoids and pick positive samples from clusters where (a) cosine distance between cluster names <= 0.08, or (b) cosine distance between cluster medoids <= 0.05 or (c) cosine distance between cluster names <= 0.1 and cosine distance between medoids <= 0.1 as positive pairs. These thresholds were tuned empirically. We sample a small number of anchor phrases with code-mixed or fully regional phrases, and we added their English translation as a positive pair to enable the model's semantic matching robustness in the presence of vernacular.

(2) Negative Pairs: If names of 2 clusters have a distance higher than a particular threshold (0.4), then the phrases from one cluster qualify to be negative pair to phrases of another cluster.

Once the text embedding model is trained and fine-tuned for e-commerce review text, we again use the same agglomerative clustering technique (as described in section 3.1) to generate robust and high quality aspect grouping. After re-clustering using the fine-tuned embeddings, we then use the T5 based aspect name generation model that was developed in section 3.2 to generate the aspect names for these new clusters. Even though the aspect name generation model wasn't re-trained in this step, but still the aspect name generation improves due to the new clusters being more coherent.

4 Experiments

4.1 Baselines

We use the following baseline algorithms to compare with our proposed framework.

(1) SBERT-STS-Clustering: We use the state-of-the-art sentence transformers (Reimers and Gurevych, 2019) model trained the STS task [1] for phrase embedding and agglomerative clustering to create aspect groups. We use this baseline to compare with our aspect grouping model that uses distant supervision.

(2) DS-Clustering: This is our proposed final clustering model as described in section 3.3.

(3) Heuristic-Name-Generation: We use the heuristic algorithm (used for distant supervision) using TFIDF scores and distance threshold as described in section 3.2 as a baseline for aspect name generation.

(4) DS-BART-Name-Generation: We train the state-of-the-art conditional language generation model, BART (Lewis et al., 2019) with our distant supervision strategy as a baseline for aspect name generation model.

(5) DS-T5-Name-Generation: This is our proposed final aspect name generation model as described in section 3.3 using T5 (Raffel et al., 2019).

4.2 Experimental Setup

We use the sentence-transformers [2], HuggingFace [3] and Pytorch [4] libraries to train our reviews phrase embedding model. Training was done on a single Nvidia V100 GPU. Batch size was set to be 16. Learning rate was set to be $2X10^{-05}$ with 10% of total training iterations as warmup steps and a linear decay schedule. We used the ADAM optimizer with parameters (beta1: 0.9, beta2: 0.999, epsilon: 10^{-8}). We train the phrase embedding model for 10 epochs. We use the python SKLearn library for agglomerative clustering. For DS-Clustering, we used a cosine distance margin of 0.5. For the baseline SBERT-STS-Clustering, we use the SBERT-STS model for phrase embedding and the agglomerative clustering threshold was set to 0.2. We train the T5 model using HuggingFace and Pytorch libraries for our aspect name generation model, DS-T5-Name-Generation. Batch size was set to be 2. Learning rate was set to be $5X10^{-05}$. We used the ADAM optimizer with parameters (beta1: 0.9, beta2: 0.999, epsilon: 10^{-8}). We train DS-T5-Name-Generation for 3 epochs. For our BART based baseline training, we set batch size to be 2, learning rate to be $5X10^{-05}$. The baseline was trained for 3 epochs. All the heuristic thresholds described in section 3 were hand-tuned experimentally.

4.3 Results

Dataset: To evaluate the proposed framework at scale, we collect customer reviews and return comments of a random sample of 1500 products

[1] https://huggingface.co/sentence-transformers/stsb-bert-base

[2] https://www.sbert.net

[3] https://huggingface.co

[4] https://pytorch.org

of a popular e-commerce service. The total number of reviews and return comments were around 40 million. These 40 million reviews/comments were broken down into review phrases. The review phrases were on an average 5.5 words long. Language of the corpus is a mix of English and common vernacular languages in India e.g. Hindi. Some phrases have mix-coded tokens from English and Hindi Language. A sentiment model was applied to remove the neutral phrases, resulting in 33 million phrases. Neutral phrases were removed in this exercise, as the intention was to understand the likes and dislikes of a customer for the product. Our goal is to cluster these phrases into coherent aspect groups and subsequently generate human readable names for these clusters.

Phrase Clustering: To evaluate aspect cluster quality, we use the popular Silhouette Score. Intuitively, it measures the closeness of samples to its own cluster as compared to other clusters. Silhouette Score computation doesn't require ground truth labels and consequently can be computed at scale. We also did a human annotation driven evaluation. We define the following two metrics: (i) intra-cluster accuracy: probability that a pair randomly selected from a cluster refers to the same aspect, and (ii) inter-cluster accuracy: probability that a pair randomly selected from different clusters refers to different aspects. We generate a random sample of intra-cluster phrase pairs and inter-cluster phrase pairs from the output of the DS-clustering and the baseline methods. The annotation team marked each pair as similar (pair belongs to same aspect) or dissimilar (pair belongs to different aspects). We estimate intra-cluster accuracy as the fraction of intra-cluster sampled pairs that were similar. Similarly, we estimate inter-cluster accuracy as the fraction of inter-cluster sampled pairs that were dissimilar. We report the clustering metrics in Table 2. Table 3 shows qualitative examples of clustering.

Table 2: Comparison of aspect clustering methods. Method A: SBER-STS-Clustering, B: DS-Clustering w/o MLM, C: DS-Clustering

	A	B	C
Silhouette Score	0.33	0.52	0.54
intra-cluster accuracy	0.88	0.88	0.91
inter-cluster accuracy	0.90	0.98	0.98

We see from table 2 that DS-Clustering improves over all baselines across all metrics. DS-Clustering improves by upto 64% over the baselines on Silhouette Score. On annotation driven inter/intra cluster accuracy, DS-Clustering is able to improve by upto 9%. DS-Clustering is able to improve over the baselines as our distantly supervised text embedding model is able to capture the unique language characteristics of e-commerce reviews where the general purpose text embedding models such as SBERT-STS fail to generalize. Examples of such cases are shown in table 3.

Table 3: Example review phrases that are correctly clustered by DS-Clustering inspite of presence of spelled errors(isound) and code mixing (paisa vasool translates to value for money). Baseline fails to cluster these.

Review Phrase	Cluster Name
isound quality is amazing	sound quality
fully paisa vasool.	value for money
truly value for each paisa spent	value for money

Aspect Name Generation: The name generation models in section 4.1 generate cluster names, which are on an average 2.7 words long. We measure the quality of the generated names by annotating 53K clusters generated by DS-Clustering. The annotation team reviewed sample phrases from each cluster and created a name that best described the aspect of the cluster as per their judgement. We treat this as ground truth and evaluate how close is the name generated via our model and the baselines to the ground truth. We measure closeness using ROUGE-F scores. The summary of metrics can be seen in table 4. We see that DS-T5-Name-Generation model outperforms both the Heuristic-Name-Generation as well as the DS-BART-Name-Generation models in all metrics showing that our model generates names that are most similar to that of human annotation team. Consequently, DS-T5-Name-Generation is able to generate human readable names using our novel distant supervision tech-

Table 4: Comparison of aspect name generation methods. Method A: Heuristic-Name-Generation, B: DS-BART-Name-Generation, C: DS-T5-Name-Generation

	A	B	C
ROUGE-1-F score	0.70	0.71	0.79
ROUGE-2-F score	0.46	0.47	0.63
ROUGE-L-F score	0.68	0.71	0.79

nique. DS-T5-Name-Generation is able to improve by 37% over Heuristic-Name-Generation on ROGUE-2 score even though distant-supervision was created through similar heuristics. This shows that the transfer learning capabilities of T5 combined with our heuristics based distant supervision results in a robust conditional NLG model without any manual labeling. Additionally, we analyzed cases where DS-T5-Name-Generation generated different names when compared to annotated names (i.e. ROUGE-L = 0) in table 5. Our model is able to perform well even in these cases. The ROUGE-L score is 0 as there is no word overlap, however, the generated names are semantically similar to the annotated names showing the semantic language understanding capabilities of our T5 based sequence to sequence model. In table 5, we report such examples. In the first example, a spelling error ("dimesions") in human annotation is leading to a Rouge score of 0, whereas our naming model generates names with correct spelling. In the second example, both the names are semantically similar.

Table 5: Examples where model generated names do not match annotated names. A: DS-T5-Naming. B: Manual Annotation

Cluster Phrases	A	B
['the dimensions too are incorrect','dimensions not appropriate for my usage']	wrong dimensions	inaccurate dimesions
['creating pain in foot', 'hurts feet on walking', 'itspainful for foot']	hurts the feet	Getting foot pain

5 Conclusion

In this paper we presented a practical aspect clustering and naming framework for e-commerce reviews. Our models leverage distant supervision thereby avoiding the need of manually labeled data. Extensive evaluations show improvement in clustering by 64% and naming by 16%. Survey results in appendix show that the approach generates more interpretable aspects when compared to an existing e-commerce baseline. We hypothesize that our novel distant supervision paradigm is generalizable across domains and in future we wish to explore the application of our novel distant supervision scheme to other domains. We also plan to explore principled approaches to handle multi-context phrases (phrase talking about multiple aspects) without needing manual annotations.

References

Wouter Bancken, Daniele Alfarone, and Jesse Davis. 2014. Automatically detecting and rating product aspects from textual customer reviews. In *Proceedings of the 1st international workshop on interactions between data mining and natural language processing at ECML/PKDD*, volume 1202, pages 1–16. CEUR-WS. org.

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 804–812.

T. Li C. Ding and W. Peng. Nonnegative matrix factorization and probabilistic latent semantic indexing: equivalence, chi-square statistic, and a hybrid method. *Proceedings of the American Association for Artificial Intelligence (AAAI), 2006.*

Lu Chen, Justin Martineau, Doreen Cheng, and Amit Sheth. 2016. Clustering for simultaneous extraction of aspects and features from reviews. In *proceedings of the 2016 conference of the north American chapter of the Association for Computational Linguistics: human language technologies*, pages 789–799.

Zhiyuan Chen, Arjun Mukherjee, and Bing Liu. 2014. Aspect extraction with automated prior knowledge learning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 347–358.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Aitor García-Pablos, Montse Cuadros, and German Rigau. 2018. W2vlda: almost unsupervised system for aspect based sentiment analysis. *Expert Systems with Applications*, 91:127–137.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 388–397.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer.

2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Arjun Mukherjee and Bing Liu. 2012. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 339–348.

Lucas Rafael Costella Pessutto, Danny Suarez Vargas, and Viviane P Moreira. 2020. Multilingual aspect clustering for sentiment analysis. *Knowledge-Based Systems*, 192:105339.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27.

Changqin Quan and Fuji Ren. 2014. Unsupervised product feature extraction for feature-oriented opinion determination. *Information Sciences*, 272:16–28.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, pages 111–120.

Danny Suarez Vargas, Lucas RC Pessutto, and Viviane Pereira Moreira. 2020. Simple unsupervised similarity-based aspect extraction. *arXiv preprint arXiv:2008.10820*.

X. Liu W. Xu and Y. Gong. Document clustering based on non-negative matrix factorization. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, 2003*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *ArXiv*, abs/1804.07461.

Zhenkai Wei, Yu Hong, Bowei Zou, Meng Cheng, and Jianmin Yao. 2020. Don't eclipse your arts due to small discrepancies: Boundary repositioning with a pointer network for aspect extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3678–3684.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2018. Double embeddings and cnn-based sequence labeling for aspect extraction. *arXiv preprint arXiv:1805.04601*.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232*.

Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia. 2010. Grouping product features using semi-supervised learning with soft-constraints. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1272–1280.

Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia. 2011. Clustering product features for opinion mining. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 347–354.

Denghui Zhang, Zixuan Yuan, Yanchi Liu, Zuohui Fu, Fuzhen Zhuang, Pengyang Wang, Haifeng Chen, and Hui Xiong. 2020. E-bert: A phrase and product knowledge enhanced language model for e-commerce. *arXiv preprint arXiv:2009.02835*.

6 Appendix

6.1 Comparison with a e-commerce baseline

We also compare the performance of our framework with the existing system at a popular e-commerce service that uses a non-negative matrix factorization (NMF) based topic modeling approach [5] on the "document-term" matrix created from the review corpus to extract aspects. A sample output of the framework is shown in table 7.The NMF based system is not able to distinguish semantically different aspects, resulting in incoherent clusters. E.g. "money, refund, wastage, value" are grouped together. Our proposed framework, howerver, is able to distinguish and capture the nuanced aspects. For example it is able to capture "value for money" as a separate aspect.

We use a human annotation driven approach to compare our proposed framework with the existing baseline. For each product type we get the aspect names generated by the topic modeling approach as well as our proposed framework. In each solution, for each aspect, we asked 3 "yes/no" questions to the annotation team.
(1) Does this aspect name describe the aspect of a product?

[5]Details can't be disclosed due to proprietary information

Table 6: User Responses to Survey. Improvement in Favorable Response quantifies how many more favorable responses were received for the DS-clustering + DS Naming framework as compared to the NMF framework.

Questions Asked	Improvement in Favorable Response
(a) Does this aspect name describe the aspect of a product?	+34.78%
(b) Is the supplementary information helping in understanding the aspect better?	+42.72%
(c) Does this help in knowing more about the customer likes and dislikes?	+42.22%

(2) Is the supplementary information helping in understanding the aspect better?
(3) Does this help in knowing more about the customer likes and dislikes?
The results of the survey is summarized in table 6. In the table, the "term" aspect refers to a cluster of reviews. "Aspect Name" refers to the name given to the cluster. "Supplementary Information" are the additional information given along with cluster and cluster name. In the case of DS-Clustering, they are a sample of review phrases belonging to the cluster. In the case of NMF Based Topic Modeling, they are the additional words obtained with each topic words. We can see the annotation team found our proposed framework to be significantly more helpful the topic modeling based baseline. Sample output of DS-Clustering + DS-T5-Name-generation is shown in table 1.

Table 7: Results NMF Based topic modeling on reviews of headphones

aspect name	related words
stopped	left, suddenly, 10, usage, earpiece, working, 15, warranty, function
money	value, waste, completely, spend, wastage, spent, want, refund, definitely
working	fine, left, speaker, button, perfectly, microphone, touch, 15, device
sound	clarity, clear, balanced, base, effect, loud, output, average, quality
range	mids, 10, meters, quite, frequency, audio, available, 500
battery	backup, hours, hrs, 10, long, drains, upto, performance, continuously