

Building Large-Scale English and Korean Datasets for Aspect-Level Sentiment Analysis in Automotive Domain

Dongmin Hyun, Junsu Cho, Hwanjo Yu*

Dept. of Computer Science and Engineering, POSTECH, Pohang, Republic of Korea
{dm.hyun, junsu7463, hwanjoyu}@postech.ac.kr

Abstract

We release large-scale datasets of users' comments in two languages, English and Korean, for aspect-level sentiment analysis in automotive domain. The datasets consist of 58,000+ comment-aspect pairs, which are the largest compared to existing datasets. In addition, this work covers new language (i.e., Korean) along with English for aspect-level sentiment analysis. We build the datasets from automotive domain to enable users (e.g., marketers in automotive companies) to analyze the voice of customers on automobiles. We also provide baseline performances for future work by evaluating recent models on the released datasets.

1 Introduction

Aspect-level sentiment analysis (ALSA) has been actively studied to understand authors' opinion on aspects from texts. For example, in a given text, "Although the *space* is smaller than most, it is the best *service* you will find in even the largest restaurants", the author's sentiment to *space* and *service* is negative and positive, respectively. Since devising deep learning models for ALSA recently received substantial attention (Zeng et al., 2019), building large-scale datasets in different languages has been an essential line of research (Rosenthal et al., 2017). However, due to the high cost of human annotation, the size and language of datasets are still limited. Specifically, only one of public datasets contains more than 20,000 instances, and existing datasets cover only three languages (i.e., English, Spanish, and Arabic).

To this end, we release large-scale datasets of users' comments with two languages such as English and Korean in automotive domain. The total size of these datasets is 58,603, which is the largest compared to existing datasets for ALSA. In addition, the datasets include new language (i.e., Korean), which extends the coverage of the datasets in terms of languages. In this work, we focus on automotive domain to build the datasets to enable users (e.g., marketers in automotive companies) to analyze the voice of customers on aspects related to automobiles.

To build the datasets, domain experts define the 12 largest automotive manufacturers (e.g., *Ford*) by production volume and their popular automobiles (e.g., *Mustang*) as aspects. Given the aspects, we collect users' comments from automotive communities in the United States and South Korea. To annotate aspect-level sentiments, we perform crowdsourcing by assigning at least three annotators to each comment-aspect pair. The annotated datasets consist of 28,571 and 30,032 comment-aspect pairs in English and Korean, respectively. Inter-annotator agreements are 0.36 (fair agreement) for English dataset and 0.54 (fair agreement) for Korean dataset in terms of Fleiss' kappa¹.

We perform extensive experiments with deep learning models for ALSA to provide the baseline performance on the released datasets for future work. The datasets are publicly available at our website.

2 Related Datasets

Researchers in ALSA have built labeled datasets as supervised learning has been a major approach to tackle ALSA. Table 1 tabulates the summary of prominent datasets for ALSA. Mitchell et al. (2013)

*Corresponding author

¹We follow Landis and Koch (1977) to interpret Fleiss' kappa.

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

Dataset	Size	Language	Domain
(Mitchell et al., 2013)	6,658 3,288	English Spanish	Organizations and persons
(Dong et al., 2014)	6,940	English	Celebrities and products
(Pontiki et al., 2014)	3,841 3,845	English English	Laptops Restaurants
(Rosenthal et al., 2017)	43,011 9,455	English Arabic	Popular events
(Wang et al., 2017)	12,587	English	UK election
(Jiang et al., 2019)	18,276	English	Restaurants
Ours	28,571 30,032	English Korean	Automobiles

Table 1: Summary of datasets built for aspect-level sentiment analysis.

annotated sentiments for aspects related to organizations and persons. Dong et al. (2014) annotated tweets that include aspects related to celebrities and products. Semantic Evaluation (SemEval) in 2014 (Pontiki et al., 2014) built two English datasets from laptop and restaurant domains. SemEval in 2017 (Rosenthal et al., 2017) built the largest dataset consisting of English tweets and a non-English dataset consisting of Arabic tweets with popular events as aspects such as named entities (e.g., *iPhone*) and geopolitical entities (e.g., *Palestine*). Wang et al. (2017) built a dataset to include multiple aspects in each text (i.e., tweet), and they define aspects related to UK election (e.g., *greens* and *labour*). Similar to Wang et al. (2017), Jiang et al. (2019) included multiple aspects in each text, and the aspects are related to restaurants (e.g., *food* and *service*). We note that a recent work built a dataset for ALSA in Korean (Song et al., 2019), but we omit it from our comparison as the dataset is not publicly available.

In this work, we release large-scale datasets consisting of users’ comments in English and Korean from automotive domain. The total size of our datasets is 58,000+, which is the largest compared to the other datasets. The datasets also include new non-English language (i.e., Korean) in addition to Spanish and Arabic. We believe that the released datasets further relieve the lack of large-scale and non-English datasets for ALSA.

3 Dataset Construction

3.1 Predefined Aspects

To cover a wide range of automotive domain, experts from Hyundai, a Korean automotive company, defined the 12 largest automotive manufacturers (e.g., *Ford*) by production volume and their popular automotive models (e.g., *Mustang*) as aspects in both English and Korean. The predefined list contains 12 automotive manufacturers and 341 automotive models. Table 2 tabulates the examples of the aspects in English.

3.2 Data Collection

We selected two online communities specializing in automobiles to collect users’ comments: Reddit³ for English comments and Bobae-Dream⁴ for Korean comments. From the online communities, we crawled about 0.3M English comments and 1M Korean comments where each comment contains at least one of the predefined aspects. We note that the number of English comments was relatively small because Reddit restricts viewing of old posts. We randomly sampled about 30,000 comments for each language for annotation.

Aspects in English		
Ford	BMW	Mustang
Toyota	Nissan	VW
Golf	M3	Fiesta
Hyundai	Corolla	Supra
Camry	Prius	Kia
Genesis	Fusion	GT-R
Taurus	Jetta	Crown
Skyline	Celica	Beetle
Passat	Explorer	Maxima

Table 2: Aspect examples

³<https://www.reddit.com/r/cars>

⁴<http://www.bobaedream.co.kr>

3.3 Annotation Using Crowdsourcing

We performed crowdsourcing to annotate a sentiment for each comment-aspect pair. For English comments, we used CrowdFlower for the annotation by following (Rosenthal et al., 2017) as in Figure 1. For Korean comments, we designed web pages for the annotation because there was no crowdsourcing company in South Korea. Annotators for English comments were English natives in CrowdFlower and annotators for Korean comments were Korean natives in POSTECH, a university in South Korea.

Annotators were asked to choose one of four choices (e.g., positive, neutral, negative, and wrong entity) for each comment-aspect pair (Figure 1). Wrong entity choice was designed to filter out comments that do not include automotive aspects. For example, *Morning*, a kind of automobiles, can be used as a general word such as “I went to a car repair shop this *morning*”. In this case, the correct choice is wrong entity as *morning* is not an automotive aspect. We also guided annotators to select neutral sentiment when a given comment-aspect pair does not belong to any other choices (e.g., positive, negative, and wrong entity). For quality control, we evaluate annotators with hidden tests, which are comment-aspect pairs annotated by us, and reject annotators who missed a large number of the tests. Each comment-aspect pair was annotated by at least three annotators.

We used the majority voting scheme to consolidate the annotations for each comment-aspect pair as done in (Rosenthal et al., 2015). The inter-annotator agreements are 0.36 (fair agreement) for English dataset and 0.54 (fair agreement) for Korean dataset in terms of Fleiss’ kappa. We speculate that the lower agreement rate for English dataset is due to the quality control being difficult because CrowdFlower allocated a large number of annotators (3,172 annotators) compared to a small number of annotators (10 annotators) for Korean dataset. Lastly, we exclude the comment-aspect pairs labeled as the wrong entity because they are irrelevant to ALSA.

3.4 Data Statistics

Table 3 shows the statistics of the annotated datasets in English and Korean⁵. The numbers of aspects after the annotation are 128 and 219 in English and Korean, respectively. We randomly divided each annotated dataset into training data (80%) and test data (20%).

4 Experiments

4.1 Experimental Settings

Evaluation Protocol We used accuracy and macro-F1, which have been major metrics for evaluating ALSA models (Li et al., 2018; Zeng et al., 2019). We randomly sampled 10% of training data as validation data. We also ran each model 10 times and reported the mean and standard deviation.

Baseline Models We selected deep learning models such as BERT-based models (i.e., AEN-BERT and LCF-BERT) and non-BERT-based models (i.e., the other models in Table 4). For the BERT-based models, we used original BERT (Devlin et al., 2019) trained on only English corpus for our English dataset. For Korean dataset, we used multilingual BERT (Devlin et al., 2019) trained on the corpus consisting of 100 languages including Korean. We omit the description for each model to save space.

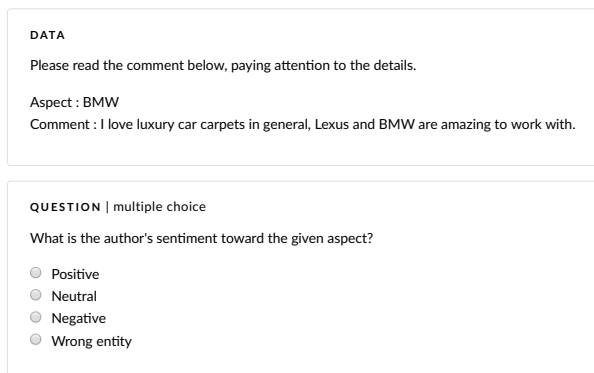


Figure 1: Annotation example for English dataset

Sentiment	English		Korean	
	Training	Test	Training	Test
Positive	7,204	1,818	4,787	1,180
Neutral	12,668	3,157	14,212	3,583
Negative	2,985	739	5,027	1,243
Total	22,857	5,714	24,026	6,006

Table 3: Statistics of datasets

⁵<https://github.com/dmhyun/alsadata>

	Dataset	English		Korean	
		Accuracy	Macro-F1	Accuracy	Macro-F1
Non-BERT	TD-LSTM (Tang et al., 2016a)	66.44 \pm 0.36	56.75 \pm 1.05	68.99 \pm 0.26	59.50 \pm 0.74
	TC-LSTM (Tang et al., 2016a)	66.97 \pm 0.33	57.16 \pm 1.06	68.66 \pm 0.46	59.13 \pm 0.86
	ATAE-LSTM (Wang et al., 2016)	66.33 \pm 0.32	56.51 \pm 1.25	67.48 \pm 0.31	56.74 \pm 1.31
	IAN (Ma et al., 2017)	66.48 \pm 0.53	55.93 \pm 1.44	63.82 \pm 0.93	47.45 \pm 2.97
	MemNet (Tang et al., 2016b)	64.17 \pm 0.21	51.48 \pm 1.81	64.75 \pm 0.22	50.89 \pm 1.01
	RAM (Chen et al., 2017)	66.83 \pm 0.34	56.67 \pm 0.87	68.68 \pm 0.32	58.01 \pm 1.00
	Cabasc (Liu et al., 2018)	63.48 \pm 0.18	50.69 \pm 0.71	65.68 \pm 0.20	52.58 \pm 0.87
	TNet-LF (Li et al., 2018)	67.55 \pm 0.27	58.43 \pm 0.95	68.84 \pm 0.45	59.84 \pm 0.84
	AOA (Huang et al., 2018)	66.80 \pm 0.37	57.57 \pm 0.77	68.90 \pm 0.30	59.15 \pm 0.73
	MGAN (Fan et al., 2018)	67.49 \pm 0.33	58.23 \pm 0.98	67.97 \pm 0.33	57.58 \pm 1.12
BERT	AEN-BERT (Song et al., 2019)	69.19 \pm 0.38	61.20 \pm 1.41	65.00 \pm 0.32	50.21 \pm 2.24
	LCF-BERT (Zeng et al., 2019)	70.72 \pm 0.48	64.09 \pm 0.86	65.30 \pm 0.55	51.17 \pm 1.77

Table 4: Classification performance of baseline models on our datasets.

Hyperparameters We used Adam (Kingma and Ba, 2014) to optimize the models using the learning rate with 0.001 for the non-BERT-based models and 0.00001 for the BERT-based models. The mini-batch size was tuned in $\{16, 32, 64, 128, 256, 1024\}$.

Word Embedding We trained Word2Vec (Mikolov et al., 2013) on the English and Korean corpora, which are crawled in this work, to obtain 100-dimensional word embedding vectors for each language, and used them to initialize words for the non-BERT-based models. In case of the BERT-based models, we used pretrained word embedding vectors included in the BERTs (i.e., original and multilingual BERT).

4.2 Performance Analysis

In Table 4, we provide the classification performance of the baseline models on the released datasets. On English dataset, the best performing model is LCF-BERT, which indicates the importance of designing ALSA models based on BERT. However, on Korean dataset, non-BERT-based models (i.e., TD-LSTM and TNet-LF) show the best performance. We speculate that the multilingual BERT is inferior to the original BERT, and investigate the performance of LCF-BERT with the multilingual BERT instead of the original BERT on English dataset. LCF-BERT with the multilingual BERT produces 64.88% of accuracy and 54.39% of macro-F1 on English dataset, which are lower than those of original LCF-BERT in Table 4. This result denotes pretraining BERT only on a target language is important to obtain better performances on the dataset in the target language. Thus, future work should pretrain BERT on large-scale Korean corpus to obtain higher performances on the released Korean dataset.

5 Conclusion

We release large-scale datasets consisting of 58,000+ comments in English and Korean from automotive domain. The total size of the datasets is currently the largest, and the datasets include new non-English (i.e., Korean) language for ALSA. For future work, we also provide the baseline performances on the released datasets using deep learning models for ALSA.

6 Acknowledgment

This research was supported by the Ministry of Science and ICT (MSIT), Korea, under the Information and Communication Technology (ICT) Consilience Creative program (IITP-2019-0-01906, IITP-2018-0-00584) supervised by the Institute for Information & communications Technology Planning & Evaluation (IITP), and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the MSIT (NRF-2020R1A2B5B03097210).

We also thank the domain experts from Hyundai, Jung-Mi Park and Kye-Yoon Kim, for helpful opinions, and Young-Woo Kim for developing Korean annotation website.

References

- Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 452–461.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *ACL (2)*, pages 49–54.
- Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3433–3442.
- Binxuan Huang, Yanglan Ou, and Kathleen M Carley. 2018. Aspect level sentiment classification with attention-over-attention neural networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 197–206. Springer.
- Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6281–6286.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.
- Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation networks for target-oriented sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 946–956.
- Qiao Liu, Haibin Zhang, Yifu Zeng, Ziqi Huang, and Zufeng Wu. 2018. Content attention model for aspect based sentiment analysis. In *Proceedings of the 2018 World Wide Web Conference*, pages 1023–1032.
- Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. *arXiv preprint arXiv:1709.00893*.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Margaret Mitchell, Jacqueline Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open domain targeted sentiment. In *EMNLP*, pages 1643–1654.
- Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August. Association for Computational Linguistics.
- Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 451–463, Denver, Colorado, June. Association for Computational Linguistics.
- Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.
- Minchae Song, Hyunjung Park, and Kyung-shik Shin. 2019. Attention-based long short-term memory network using sentiment lexicon embedding for aspect-level sentiment analysis in korean. *Information Processing & Management*, 56(3):637–653.

- Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016a. Effective lstms for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, pages 3298–3307.
- Duyu Tang, Bing Qin, and Ting Liu. 2016b. Aspect level sentiment classification with deep memory network. *arXiv preprint arXiv:1605.08900*.
- Yequan Wang, Minlie Huang, Li Zhao, and Xiaoyan Zhu. 2016. Attention-based lstm for aspect-level sentiment classification. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615.
- Bo Wang, Maria Liakata, Arkaitz Zubiaga, and Rob Procter. 2017. Tdparse-multi-target-specific sentiment recognition on twitter. pages 483–493.
- Biqing Zeng, Heng Yang, Ruyang Xu, Wu Zhou, and Xuli Han. 2019. Lcf: A local context focus mechanism for aspect-based sentiment classification. *Applied Sciences*, 9(16):3389.