Discriminative Models Can Still Outperform Generative Models in Aspect Based Sentiment Analysis

Anonymous ACL submission

Abstract

Aspect-based Sentiment Analysis (ABSA) helps to explain customers' opinions towards products and services. In the past, ABSA models were discriminative, but more recently generative models have been used to generate aspects and polarities directly from text. In contrast, discriminative models commonly first select aspects from the text, and then classify the aspect's polarity. Previous results showed that generative models outperform discriminative models on several English ABSA datasets. Here, we evaluate and contrast two state-of-the-art discriminative and generative models in several settings: cross-lingual, cross-domain, and cross-lingual and domain, to understand generalizability in settings other than English monolingual in-domain. Our more thorough evaluation shows that, contrary to previous studies, discriminative models can still outperform generative models in almost all settings.

1 Introduction

Online reviews make it easy for customers to share their feelings about products and services in a quick and efficient way. But for business owners, this can mean a deluge of comments with a variety of concerns. Companies with millions of customers receive massive amounts of online reviews that can't be analyzed manually, thus needing automation.

Some natural languages receive more research effort compared to other languages (e.g. English vs. Swahili). Although the community has remarkably accelerated the improvement of English NLP techniques, techniques for other languages lag behind. Working on a lower resource language is a challenging task, where few datasets, lexicons, and models exist. Thus, utilizing cross-lingual approaches is important to migrate knowledge across languages.

In low resource settings, it can be difficult to use techniques like ABSA to analyze reviews. ABSA involves predicting aspect terms and their associated sentiment polarities (Liu, 2012), which requires a fair amount of training data. For example, "Service was good at the restaurant, but food was not" has two aspect terms ("service" and "food"), associated with sentiments "positive" and "negative", respectively. In this work, we conduct a comparative study of two ABSA model types (discriminative and generative). Discriminative models, which use decision boundaries to make predictions, commonly use sequence labeling techniques to detect aspects in a given review (extraction) and then use another step to classify those aspects (classification). On the other hand, generative models use encoder-decoder language models to learn probability distributions, and generate aspects and sentiment polarities together without separate extraction and classification steps. Notably, a few discriminative models do extraction and classification at once (Li et al., 2020, 2019a; Hu et al., 2019). However, results showed that training both tasks together does not always improve performance.

Previous work has shown that generative models achieve better performance than discriminative models when trained and evaluated on the English in-domain setting (Zhang et al., 2021; Yan et al., 2021). While recent studies compared generative to discriminative models in English in-domain setting, none have explored performance in cross-lingual or cross-domain settings. Here, we evaluate the performance of the two model types in cross-lingual and cross-domain settings by comparing the state-of-the-art representatives. Additionally, we propose a more challenging setting: both cross-lingual and cross-domain. Our results demonstrate that discriminative models can still perform better than generative models in almost all proposed scenarios.

2 Methodology and Experimental Setup

2.1 Datasets

In our experiments, we consider several languages and domains for a more valid evaluation. For

Datasets	Data Split	#Pos	#Neg	#Neu
	Train	864	313	47
Rest16$_{en}$	Val	130	32	6
	Test	427	119	28
	Train	972	338	72
Rest16$_{es}$	Val	101	46	5
	Test	420	142	29
	Train	1068	216	99
Rest16$_{ru}$	Val	223	56	23
	Test	608	193	85
	Train	591	515	268
Lap14	Val	99	71	50
	Test	341	128	169
	Train	636	552	982
MAMS$_{En}$	Val	403	325	605
	Test	400	330	607

Table 1: Datasets' statistics - Count of aspects with sentiment polarities for the sampled and cleaned datasets. Multiple aspects can exist in single record

languages we use SemEval datasets - Restaurant (Rest16) (Pontiki et al., 2016) in English, Spanish and Russian. For domains we use Rest16 and Laptop (Lap14) from SemEval (Pontiki et al., 2014) which are widely used in the literature (Li et al., 2019b; Tian et al., 2021; Liang et al., 2021). As an additional domain, we use MAMS dataset for ABSA (Jiang et al., 2019). MAMS (Jiang et al., 2019) is a recently developed challenge dataset in which each sentence contains at least two aspects with different polarities, making the dataset more challenging than the SemEval datasets.

For SemEval datasets, since the validation sets are not given, we sample 10% of the training dataset to use for validation. The datasets we considered vary in terms of the type of content and the training set size. Thus, for a fair comparison, we reduce the larger training datasets to have an equal number of records. For this purpose, we sample without replacement 857 records from each training dataset, which is the minimum number of training instances across datasets (cleaned Rest16$_{es}$ training dataset has 857 records). Table 1 presents the datasets' statistics after cleaning and sampling. A larger dataset could improve model performance, but we must control for dataset size to ensure fair comparisons across datasets. For example, training on a larger datasets like MAMS may give better performance when testing across domain on Rest16$_{en}$ than if we had trained and tested in domain with Rest16$_{en}$. In that case, we would not be able to determine if the effect was due to dataset size, or some inter-domain interaction.

2.2 Models and Baselines

For contrasting generative and discriminative model types, we consider a representative model for each, which shows state of the art performance in the ABSA task. For the generative model, we use the approach proposed in (Zhang et al., 2021), which is an encoder-decoder T5-based model. This model takes a review as input and generates the aspects with their polarities. The aspect-polarity terms have the following format: "service positive <sep> food negative", indicating the presence of two aspect terms ("service" and "food"), with the associated polarities ("positive" and "negative"). Since there can be multiple aspect-polarity pairs in a single review, we add a separator token "<sep>" to demarcate a separation between multiple aspect-polarity pairs. In the mono-lingual setting, the model is trained on English and generates English aspect-polarity pairs. When we move to the cross-lingual setting, we ask a multilingual model to generate aspect-polarity pairs for a language that was not used in the training process. Thus, we use an approach that augments the training data with a version of itself translated automatically to the test language (Riabi et al., 2021). This does not require additional annotated data to solve the issue. In Appendix A.2, we give more details regarding this approach.

For the discriminative model, we consider the SPAN-BERT model (Hu et al., 2019) which is a state-of-the-art model that uses a BERT transformer. It has a good performance in mono-lingual in-domain datasets, and has been used as a baseline for the generative model released by Zhang et al. (2021). The model extracts spans (continuous spans of text) for multiple target aspect terms using a decoder heuristic and then classifies their polarities using contextualised span representations. The discriminative and generative models referenced above use transformers trained solely on English, so we need to modify them before training on other languages. To make our experiments consistent, we use multilingual versions of the base transformers. For the generative model, we use the multilingual T5 (mT5-base) model (huggingface implementation[1]). For the SPAN-BERT model, we use the multilingual BERT model from Google[2].

[1]https://huggingface.co/transformers/model_doc/mt5.html

[2]https://github.com/google-research/bert/blob/master/multilingual.md

Domain$_{Lang}$	Discriminative	Generative
Rest16$_{En}$	0.56	**0.58**
Rest16$_{Es}$	**0.63**	0.58
Rest16$_{Ru}$	**0.47**	0.42
Lap14$_{En}$	**0.50**	0.36
MAMS$_{En}$	**0.54**	0.44

Table 2: Mono-lingual and in-domain F1 scores. Bolded results are the best among models.

Train → Test	Discriminative	Generative
$Es \rightarrow En$	0.51 (-6%)	0.34 (-24%)
$Ru \rightarrow En$	**0.53** (-3%)	**0.45 (-13%)**
$En \rightarrow Ru$	**0.44** (-3%)	0.27 (-15%)
$Es \rightarrow Ru$	0.42 (-5%)	**0.29 (-13%)**
$En \rightarrow Es$	**0.54** (-9%)	0.39 (-19%)
$Ru \rightarrow Es$	0.52 (-11%)	**0.45 (-13%)**

Table 3: Cross-lingual F1 scores using Rest16 in several languages. Bolded results are best per model and test language. Bracketed %s show performance decrease compared to the mono-lingual, in-domain result 2.

Train → Test	Discriminative	Generative
Rest16$_{En}$ → Lap14$_{En}$	0.29 (-21%)	**0.21** (-15%)
MAMS$_{En}$ → Lap14$_{En}$	**0.31** (-19%)	0.19 (-17%)
Lap14$_{En}$ → Rest16$_{En}$	0.44 (-12%)	0.21 (-37%)
MAMS$_{En}$ → Rest16$_{En}$	**0.47** (-9%)	**0.38** (-20%)
Rest16$_{En}$ → MAMS$_{En}$	**0.32** (-22%)	**0.3** (-14%)
Lap14$_{En}$ → MAMS$_{En}$	0.29 (-25%)	0.12 (-32%)

Table 4: Cross-domain F1 scores. Bolded results are the best per model and test language. Bracketed % values show performance decrease compared to the mono-lingual, in-domain result 2.

In order to understand the performance of both models, we set two baselines: mono-lingual in-domain, and a random selection baseline. In the mono-lingual in-domain, we train each model on each dataset to define the theoretical performance ceiling. The random baseline will allow us to see if our cross-lingual or domain results are better than chance. In the random baseline, we have the model pick aspect words from the text (excluding stop words), and their polarities at random. For further details refer to Appendix A.5. For consistency of results, we apply the same data normalization steps for both models. Appendix A.3 gives details of the normalization and matching process.

3 Results and Discussion

3.1 Monolingual and In-Domain

First, we evaluate models with the train and test data of the same dataset type and language, and we get the results of the random selection baseline. Table 2 presents the results. For detailed results refer to Appendix A.6. From a mono-lingual perspective, we can see that the discriminative model performs better than the generative model for all datasets except *Rest16$_{en}$*. During our experiments, we also evaluated models using the mono-lingual version of the transformers models, and we had noticed a similar scenario; the generative approach performed better than the discriminative one in only Rest16$_{en}$ and Lap14$_{en}$ datasets. Thus, it seems that the generative approach can only perform better on English datasets. The random baseline results in all the datasets are around 4% F1 (individual results can be seen in 6)

3.2 Cross-Lingual

Table 3 presents the cross-lingual results. For detailed results refer to Appendix A.6. From a cross-lingual perspective, we can clearly see that all models, perform above random guess which is nearly 4% F1 (Table 6). Regarding the discriminative model with the Spanish and Russian test sets, we notice that we obtain the highest results when we train on English. And the largest decrease in performance is when we train on Russian and test on Spanish. Interestingly, when we train on Russian and test on the other languages, we obtain the highest results for the generative model. Overall, the performance drop of the generative cross-lingual results compared to the monolingual ones is high, considering the discriminative model's results. We can conclude that the discriminative model generalizes better than the generative one in the cross-lingual setting.

3.3 Cross-Domain

Table 4 presents the cross-domain results. More details can be found in Appendix A.6. Generally, considering both models' results, training on Rest16$_{En}$ and MAMS$_{En}$ datasets produced the highest results. Like the Rest16 dataset, MAMS contains reviews related to restaurants. Thus it is not surprising that training on one of these two datasets and testing on the other gives higher results compared to training on Lap14. However, we can see that this gap is larger when we experiment with the generative model. This observation demonstrates that the generative model is more domain sensitive.

3.4 Cross-Lingual and Cross-Domain

In this experiment, we evaluate both models in an extreme setting, which combines the previous cross-lingual and cross-domain. Table 5 shows the

Train → Test	Discriminative	Generative
Rest16$_{Es}$ → Lap14$_{En}$	**0.3** (-20%)	**0.17** (-19%)
Rest16$_{Ru}$ → Lap14$_{En}$	0.28 (-22%)	0.16 (-20%)
Lap14$_{En}$ → Rest16$_{Es}$	0.54 (-9%)	0.33 (-25%)
Lap14$_{En}$ → Rest16$_{Ru}$	0.34 (-13%)	0.27 (-15%)

Table 5: Cross-domain and cross-lingual F1 scores. Bolded results are the best per model and test language, when more than 1 train language to compare. Bracketed % values show performance decrease compared to the mono-lingual, in-domain result 2.

evaluation results. More details can be found in Appendix A.6. We can see a larger drop compared to the results in the cross-lingual experiment (see 3), except when we test on Rest16$_{es}$ using the discriminative model; training on Rest16$_{en}$ or Lap14$_{en}$ gives the same F1 result. Similar to the previous results, the generative model achieves lower results compared to the discriminative one.

4 Discussions and Conclusion

In this work, we compared two types of ABSA models in terms of performance differences by considering a state-of-the-art model for each type as a representative. We compared those models across languages and domains. Previous studies showed that generative models achieve higher results than the discriminative ones across almost all the available English ABSA datasets. However, the results in our study demonstrated that generative models can perform worse than the discriminative ones in almost all of the proposed scenarios, namely, cross-lingual, cross-domain, and cross- lingual and domain. We conduct an error analysis to understand the scenarios where the models fail in Appendix A.4.

We experimented with datasets from three languages, and from three different domains. Briefly, the results showed that generative models can be more language and domain sensitive. Generative models have a challenging task: they must learn a joint probability over all words. This is in contrast to discriminative models which need only learn a small set of decision boundaries. Generative models sample words from the entire data distribution and might be more sensitive to the training data size compared to discriminative models which classify only the words in the original sentence. This intuition is supported by existing literature (Ng and Jordan, 2002). Given that we have only 857 instances for training, the generative model did not generalize as good as the discriminative in the other domains or languages. Additionally, it is possible that the evaluation process is very strict and hurts the generative model (see Appendix A.3).

The generative model outperformed the discriminative model in only one English mono-lingual experiment, perhaps due to a favourable bias in the mT5 model towards the English language. Recent studies showed that Multilingual encoder-decoder transformers do not perform well in languages other than English (Tang et al., 2020; Fan et al., 2021). Notably, we see that the discriminative model does better on English monolingual in-domain cases of Lap14$_{en}$ (contrary to Zhang et al. (2021)) and MAMS$_{en}$. We attribute this apparent contradiction for Lap14$_{en}$ to the different evaluation technique and the multilingual encoder variant. Moreover, in case of MAMS$_{en}$, a small sized training dataset could have had adverse effects since it is a challenging dataset.

Another reason for the variation in results could be that each model uses a different encoder. The discriminative model uses mBERT encoder whereas the generative one uses an mT5 encoder. We do not make the encoders consistent in the models to avoid making drastic changes to ABSA models with proven good performance in the literature. Nevertheless, our results are a useful comparison of generative and discriminative model types since we compare their state-of-the-art representatives to draw conclusions on why generative models are not always preferable.

Considering the random selection baseline in our experiments, we can conclude that generative models are capable of generating correct aspects and polarities. The results showed that the generative model, in the worst case (training on Lap14$_{En}$ and testing on MAMS$_{En}$), performs better than the random baseline by 8% F1. On the other hand, the discriminative model in the worst case (training on Rest16$_{Ru}$ and testing on Lap14$_{En}$), performed better than the random baseline by 25% F1.

These results argue against adopting generative models as the defacto standard for all ABSA tasks as discriminative models are more accurate in some settings. For future work, we plan to study other generative models in this task. We also plan to study both types of models in other scenarios like conflicting polarities (aspects with both positive and negative polarities).

References

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22:1–48.

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-domain targeted sentiment analysis via span-based extraction and classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 537–546, Florence, Italy. Association for Computational Linguistics.

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 6280–6285, Hong Kong, China. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Jiawen Li, Yudianto Sujana, and Hung-Yu Kao. 2020. Exploiting microblog conversation structures to detect rumors. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5420–5429, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019a. A unified model for opinion target extraction and target sentiment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6714–6721.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019b. Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 34–41, Hong Kong, China. Association for Computational Linguistics.

Yunlong Liang, Fandong Meng, Jinchao Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. A dependency syntactic knowledge augmented interactive architecture for end-to-end aspect-based sentiment analysis. *Neurocomputing*, 454:291–302.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Andrew Y Ng and Michael I Jordan. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. 2021. Synthetic data augmentation for zero-shot cross-lingual question answering. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224, Austin, Texas. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Yuanhe Tian, Guimin Chen, and Yan Song. 2021. Aspect-based sentiment analysis with type-aware graph convolutional networks and layer ensemble. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2910–2922, Online. Association for Computational Linguistics.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2416–2429, Online. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational

Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 504–510, Online. Association for Computational Linguistics.

A Appendix

A.1 Dataset Cleanup

We follow existing work (Tian et al., 2021; Tang et al., 2016) in removing sentences with no opinions (not useful for the considered ABSA task), as well as sentences having aspect terms with a "conflict" sentiment polarity, from the dataset. This is to prevent a class imbalance problem, as there are very few instances of "conflict", compared to other polarities.

A.2 Generative models with Cross-lingual Setting

In this section, we provide more details regarding the proposed approach in (Riabi et al., 2021) to solve the issue of controlling the generated language. The idea of the method is that, for instance, when we train on English and generate for Spanish, we translate the English training data to Spanish (using Google Translator) and we include it in the training part with the original English language. Additionally, to control the target language, we use a specific prompt (token) per language (<LANG>), which corresponds to the desired target language (e.g. Spanish : Spanish_review). When we translate a language into another, we discard instances that their translated aspect terms do not exist in the translated review. This is important for SPAN-BERT models as terms indices are needed. Also, we sample an equal number of translated training instances in all the languages (507 instances per language), as we prepared the monolingual training data. For consistency, we train SPAN-BERT model on the same data.

A.3 Preprocessing for Evaluation

We find that the generative model sometimes generates a different variant of a term, e.g. plural or singular. Prior to evaluating the model outputs, we perform a normalisation process. For normalising, we remove punctuation marks such as ",", ".", "“”" from the sentences, lower-case and lemmatise the words, and remove common stop words. This idea of normalising the generated output is similar to Zhang et al. (2021), where Levenshtein distance is used to align the generated aspect words with the closest words existing in the original sentence. Compared to this, our normalisation process followed by an exact matching is stricter. Levenshtein distance may align the model's predictions with unrelated words in the original sentence. For example, if a generated word - "salmon", has the least distance with the word "not" out of all the words in the original sentence, then "salmon" can get aligned to "not", as is mentioned by Zhang et al. (2021), which is a loose matching.

After model outputs and gold data are normalised, an exact match is made to compare the predicted aspect polarity terms with corresponding aspect polarity terms in the gold data. We consider a hit only if both the aspect term and the polarity term match. We use the standard evaluation metrics for calculating ABSA scores, which are Micro-Precision, Recall and F1. We use the evaluation code released by Li et al. (2019a)[3].

A.4 Error Analysis

We conduct an error analysis on the outputs of the models to better understand the cases where they fail.

For the discriminative model, we found that in a large number of the error cases (nearly 40% in Rest_{en} monolingual in-domain case), the model did not predict any aspect term at all. This means that the SpanBert model was not able to confidently identify any possible aspect term spans. E.g. for the following sentence, the model fails to predict an aspect term: "Not the biggest portions but adequate." Since SpanBert uses thresholds to compute scores (representing confidence), it is possible that the model was not sufficiently confident in its predictions. we also found several cases where the model gets the aspect term correct but the sentiment incorrect, such as for the sentence "i am never disappointed with there food."; it gives "food" a negative sentiment instead of a positive. Here, it can be seen that the underlying language model (BERT) did not understand the word "never" in the sentence, and instead understood the sentiment from "disappointed" which has negative connotations. This can be attributed to the fact that language models like BERT misunderstand some negations (Kassner and Schütze, 2020). Another common error is where the predicted and gold spans have some overlap, but are not an exact match. This can be seen in cases such as "La atención del personal impecable." ("The attention of the impeccable staff.") where the

[3] http://github.com/lixin4ever/E2E-TBSA

predicted aspect term is "personal" ("staff") instead of "atención del personal" ("attention of the staff").

As in the discriminative model, in the generative model we saw several cases where the predicted phrase is incorrect, though it refers to the same entity conceptually. For instance, in the sentence "Great draft and bottle selection and the pizza rocks.", the predicted entities include "bottle selection" instead of "draft and bottle selection". Such predictions would not have been considered errors if we had gone with a partial matching approach like Zhang et al. (2021). Other notable cases included those where a similar entity is predicted instead. For example, for the sentence - "The best calamari in Seattle!", the generative model generated "salmon" as an aspect term instead of "calamari". This does show that the language model understood the similarity between calamari and salmon, however it did not understand that for the task it was supposed to predict a word from the input sentence itself, and not make such inferences. Similarly, for the sentence "Un sitio recomendable en pleno centro de barcelona" ("A recommended place in the heart of barcelona"), it generates "comida" ("meal") as an aspect term instead of "sitio" ("site"). There is also an indication the model needs more data to understand adjectives. This can be seen from examples such as - for the sentence "Mediocre food", the model predicts "food" as positive instead of neutral as it is misinterprets "Mediocre".

A.5 Random Baseline

We consider a random model for evaluating the performance of the considered models. However, we do not simply randomly assign positive, negative, neutral or none labels to randomly selected words in a sentence. Instead, we produce predictions in a slightly less strict way. Firstly, we select aspect words from words in the sentences (disregarding stop words). Then, we consider the distribution of different polarities for the aspect terms in the gold predictions (this gives our random baseline a positively biased edge), and assign polarities to the aspect words in the training data based on those distributions.

A.6 Detailed Results

In this section we have the detailed results for the experiments we conducted. The precision, recall and F1 values can be found here.

Table 6 gives the detailed results for the experiments conducted for the mono-lingual in-domain case, including the results for the random baseline. This contains more details compared to Table 2. Similarly, we have Tables 7, 8 and 9 which are detailed versions of Tables 3, 4 and 5 respectively

Domain$_{Lang}$	**Discriminative**			**Generative**			**Random Selection**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Rest16$_{En}$	0.67	0.48	0.56	0.64	0.52	**0.58**	0.07	0.04	0.05
Rest16$_{Es}$	0.65	0.60	**0.63**	0.67	0.51	0.58	0.07	0.03	0.05
Rest16$_{Ru}$	0.47	0.48	**0.47**	0.46	0.39	0.42	0.06	0.04	0.05
Lap14$_{En}$	0.48	0.52	**0.50**	0.4	0.33	0.36	0.05	0.02	0.03
MAMS$_{En}$	0.53	0.55	**0.54**	0.48	0.4	0.44	0.06	0.03	0.04

Table 6: Mono-lingual and in-domain results. Bolded results are the best among models.

Train $\rightarrow$ **Test**	**Discriminative**			**Generative**		
	P	**R**	**F1**	**P**	**R**	**F1**
Rest16$_{Es}$ $\rightarrow$ **Rest16**$_{En}$	0.58	0.45	0.51 (-6%)	0.48	0.26	0.34 (-24%)
Rest16$_{Ru}$ $\rightarrow$ **Rest16**$_{En}$	0.55	0.51	**0.53** (-3%)	0.6	0.36	**0.45 (-13%)**
Rest16$_{En}$ $\rightarrow$ **Rest16**$_{Ru}$	0.53	0.37	**0.44** (-3%)	0.43	0.20	0.27 (-15%)
Rest16$_{Es}$ $\rightarrow$ **Rest16**$_{Ru}$	0.42	0.43	0.42 (-5%)	0.52	0.21	**0.29 (-13%)**
Rest16$_{En}$ $\rightarrow$ **Rest16**$_{Es}$	0.75	0.42	**0.54** (-9%)	0.55	0.3	0.39 (-19%)
Rest16$_{Ru}$ $\rightarrow$ **Rest16**$_{Es}$	0.59	0.46	0.52 (-11%)	0.62	0.35	**0.45 (-13%)**

Table 7: Cross-lingual results. Bolded results are the best per model and test language. The percentage values between brackets represent the amount of drop compared to the mono-lingual and in-domain result.

Train $\rightarrow$ **Test**	**Discriminative**			**Generative**		
	P	**R**	**F1**	**P**	**R**	**F1**
Rest16$_{En}$ $\rightarrow$ **Lap14**$_{En}$	0.28	0.3	0.29 (-21%)	0.42	0.14	**0.21** (-15%)
MAMS$_{En}$ $\rightarrow$ **Lap14**$_{En}$	0.41	0.25	**0.31** (-19%)	0.23	0.16	0.19 (-17%)
Lap14$_{En}$ $\rightarrow$ **Rest16**$_{En}$	0.46	0.43	0.44 (-12%)	0.34	0.15	0.21 (-37%)
MAMS$_{En}$ $\rightarrow$ **Rest16**$_{En}$	0.51	0.44	**0.47** (-9%)	0.36	0.42	**0.38** (-20%)
Rest16$_{En}$ $\rightarrow$ **MAMS**$_{En}$	0.38	0.27	**0.32** (-22%)	0.39	0.24	**0.3** (-14%)
Lap14$_{En}$ $\rightarrow$ **MAMS**$_{En}$	0.33	0.27	0.29 (-25%)	0.29	0.07	0.12 (-32%)

Table 8: Cross-domain results. Bolded results are the best per model and test language. The percentage values between brackets represent the amount of drop compared to the mono-lingual and in-domain result.

Train $\rightarrow$ **Test**	**Discriminative**			**Generative**		
	P	**R**	**F1**	**P**	**R**	**F1**
Rest16$_{Es}$ $\rightarrow$ **Lap14**$_{En}$	0.31	0.28	**0.3** (-20%)	0.31	0.11	**0.17** (-19%)
Rest16$_{Ru}$ $\rightarrow$ **Lap14**$_{En}$	0.3	0.26	0.28 (-22%)	0.24	0.12	0.16 (-20%)
Lap14$_{En}$ $\rightarrow$ **Rest16**$_{Es}$	0.53	0.56	**0.54** (-9%)	0.48	0.25	**0.33** (-25%)
Lap14$_{En}$ $\rightarrow$ **Rest16**$_{Ru}$	0.53	0.25	0.34 (-13%)	0.47	0.18	0.27 (-15%)

Table 9: Cross-domain and cross-lingual results. Bolded results are the best per model and test language. The percentage values between brackets represent the amount of drop compared to the mono-lingual and in-domain result.