Aspect Based Sentiment Analysis into the Wild

Caroline Brun, Vassilina Nikoulina*
firstname.lastname@naverlabs.com
Naver Labs Europe, 6 chemin de Maupertuis, 38240 Meylan

Abstract

In this paper, we test state-of-the-art Aspect Based Sentiment Analysis (ABSA) systems trained on a widely used dataset on actual data. We created a new manually annotated dataset of user generated data from the same domain as the training dataset, but from other sources and analyse the differences between the new and the standard ABSA dataset. We then analyse the results in performance of different versions of the same system on both datasets. We also propose light adaptation methods to increase system robustness.

1 Introduction

The aim of Aspect Based Sentiment Analysis (ABSA) is to detect fine-grained opinions expressed about different aspects of a given entity, on user-generated comments.

Aspects are attributes of an entity, e.g. the screen of a cell phone, the service for a restaurant, or the picture quality of a camera, and can be described by an ontology associated to the entity. ABSA includes therefore to identify aspects of an entity, and the sentiment expressed by the writer of the comment about different aspects. For example, from a sentence extracted from a review about a museum, an ABSA system could extract the following information: *This museum hosts remarkable collections, however, prices are quite high and the attendants are not always friendly.*

”collections”: aspect=museum#collection, polarity=positive;
”prices”: aspect=museum#price, polarity=negative;
”attendants”: aspect=museum#service, polarity=negative;

ABSA receives now a specific interest from the scientific community, especially with the SemEval dedicated challenges, (Pontiki et al., 2014), (Pontiki et al., 2015), (Pontiki et al., 2016), that provided a framework to design and evaluate ABSA systems, for different domains, initially on English but for 8 languages in the 2016 (last) edition. Besides SemEval, other challenges focussing on the task have been also launched recently, for example TASS, dedicated to Spanish, (Villena-Román et al., 2015b), (Villena-Román et al., 2015a), (Cumbreras et al., 2016), or GermanEval, dedicated to German ABSA, (Wojatzki et al., 2017).

Following this particular interest, the technology performing ABSA becomes more and more mature, however, experiments and evaluation are restricted to a small number of academic datasets, in relatively favorable settings. The goal of this paper is to test a state-of-the-art ABSA system on actual data, to evaluate the performance loss in real-world application conditions, and to experiment potential solutions to it. To achieve this goal, we've created a new ABSA annotated dataset, developed on Foursquare data. We also performed evaluation of the full ABSA processing chain (as opposed to sub-tasks evaluation which is traditionally performed). We also propose a weakly supervised method for aspect-based lexical acquisition designed to improve the robustness of our initial system.

2 Related Work

Most of the systems dedicated to ABSA use machine learning algorithms such as SVMs (Wagner et al., 2014; Kiritchenko et al., 2014), or CRFs (Toh and Wang, 2014; Hamdan et al., 2015), which are often combined with semantic lexical information, n-gram models, and sometimes more fine-grained syntactic or semantic information. For example, (Kumar et al., 2016) proposed a very efficient system on different languages of SemEval2016. The system use information extracted from dependency graphs and distributional thesaurus learned on the different domains and

*Both authors contributed equally.

Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 116–122

https://doi.org/10.18653/v1/P17

languages of the challenge. Deep Learning methods are also emerging: for example, (Ruder et al., 2016) proposed a method using multiple filters CNNs and obtained competitive results on both polarity and aspect detection tasks. However, ABSA datasets are very costly to annotate by humans, and they are usually small, which is a problem for Deep Learning supervised methods.

3 Datasets

Usually, ABSA systems are tested on the same dataset as they are developped on. One of the widely used ABSA datasets was released in Semeval2016 challenge (Pontiki et al., 2016), in particular the dataset for restaurant domain. It is based on the dataset of (Ganu et al., 2009) who extracted restaurant reviews from City Search New York over year 2006. Since then, the notion of the user review has evolved. Many factors may impact the linguistic structure of a review, e.g. the support it was written on (computer vs. smartphone), the age of the user, the location (US vs. UK English), the user mother tongue (native vs. non-native speakers), etc. How would a system trained on Semeval2016 dataset perform on a new data coming from different sources?

In order to assess ABSA real-world performances, we manually annotated a completely new dataset from Foursquare[1] comments. We have access to about 215K user reviews of restaurants all over the world in English[2]. The reviews were written during the period between 2009 to 2018. From these reviews, we randomly selected 585 samples, which contain 1006 sentences and annotate these sentences with the SemEval2016 annotation guidelines for the restaurant domain. The annotations have been performed by a single annotator, expert linguist with a very good knowledge of the SemEval2016 annotation guidelines, using BRAT, (Stenetorp et al., 2012).

Each sentence contains annotations about: 1. Opinion Target Expression (OTE), i.e. the linguistic expression (term) used in the text to refer to the reviewed entity, annotated as "NULL" if the aspect is implicit; 2. Aspect Categories, i.e. the semantic categories of the opinionated aspects, which are part of a predefined ontology (12 semantic classes for the restaurant domain from (Pontiki et al., 2016)); 3. Sentiment Polarities: polarities (positive, negative or neutral) associated to the tuple <OTE, Aspect Category>. An illustration of such annotation is given on figure 1.

Dataset	#Rev	#S	#W/S	#A/S
Semeval	92	676	12.8	1.27
Foursquare	585	1006	8.0	1.15

Table 1: Dataset statistics: Semeval 2016 test set and Foursquare dataset. #Rev: number of reviews, #S: number of sentences, #W/S : number of words per sentence, #A/S: number of <OTE, Aspect Category, Polarity> tuples per sentence

```
<text>Their sake list was extensive,
but we were looking for Purple Haze,
which wasn't listed but made for us
upon request!</text>
<Opinions>
<Opinion target="sake list"
         category="DRINKS#STYLE_OPTIONS"
         polarity="positive"/>
<Opinion target="NULL"
         category="SERVICE#GENERAL"
         polarity="positive"/>
</Opinions>
```

Figure 1: ABSA: an annotated sentence from the Semeval-2016 training corpus

Table 1 gives some statistics about the Foursquare and Semeval2016 datasets. One may notice, that in average, Foursquare reviews are shorter and therefore contain less aspects per sentence. We believe this is due to the generalisation of smart-phones (and other mobile devices) usage over the world in the last decade, which influenced the way users write. We release the Foursquare dataset to the community in order to better assess robustness of ABSA systems[3].

4 Evaluation Procedure

We consider different evaluation measures. First, we re-use the SemEval2016 ABSA evaluation paradigm and scripts, where the evaluation was run in two phases, phase A and phase B. In phase A, raw reviews have to be annotated with aspects (slot 1 of the challenge) and OTE (slot 2 of the challenge). In phase B, gold annotations for phase A, i.e. tuples <OTE, aspect>, have to be annotated with polarities (slot 3 of the challenge).

[1] https://foursquare.com/

[2] Countries with most of English comments include US, UK, Australia, Canada, Indonesia, Malaysia, Philippines, India, Thailand

[3] http://www.europe.naverlabs.com/Research/Natural-Language-Processing/Aspect-Based-Sentiment-Analysis-Dataset

Thus, we evaluate separately the OTE detection, aspect detection and finally, we evaluate the polarity of opinion detection on the ground truth of phase A. The advantage of this evaluation procedure is of course to assess the quality of the systems on each of the different subtasks involved in the full ABSA system. However, these measures do not reflect the overall results such systems would obtain on the full chain of annotations starting from raw data, in end-to-end application settings. Therefore, we also propose to evaluate the results obtained with the complete annotation chain, i.e. computing F1-measure on the triplets <OTE, Aspect, Polarity>. In addition, we compute the F1-measure on the pairs <Aspect, Polarity> at sentence level. This last measure can be useful to assess ABSA general Aspect-Polarity performance since many ABSA applications may not require the OTE step. In what follows, we refer to these measures as slot1,3 and slot1,2,3 to make connection with the challenge tasks.

5 Baseline ABSA Systems

In our experiments, we use several baseline systems. Each of the systems consists in the following pipeline of different components: 1. Opinionated domain term extraction (OTE); 2. aspect categorization, for opinionated term (OT), and whole sentence level; final aspect is predicted as a combination of both; 3. polarity classification of each aspect identified in the previous step. The difference between baselines lies in the implementation of each component of the pipeline, and the level of external resources involved.

5.1 Baseline-1

The first system is resource-rich system relying on available syntactic and semantic parser, and domain-specific semantic lexicons. It is based on composite models combining sophisticated linguistic features with machine learning algorithms. The linguistic features are extracted via a NLP pipeline (based on in-house parser) comprising lexical semantic information, POS tagging, syntactic parsing and a partial semantic parsing that outputs semantic relations between polarity predicates and their opinionated targets (OTE). These linguistic features are then used by classifiers to perform each step of the pipeline.

The OTE detection is performed with Conditional Random Fields (implemented with CRF++[4] toolkit), trained with some standard features (POS, lemma, presence of upper-case letters, features combining syntactic/semantic dependencies with semantic lexicons, embedding-based features).

Aspect and polarity classification components rely on the same features as for OTE, excluding embedding-based features, but extended with bi-grams features. In addition, polarity classifier feature representation is extended with entity and attribute of aspect category (e.g. RESTAURANT#PRICES results in two additional features: *(restaurant, prices)*). Classification is performed with CoreNLP (Manning et al., 2014) implementation of Maximum Entropy.

5.2 Baseline-2

The second baseline system (*baseline-2*) replaces each component of the previous pipeline with neural network classifiers. Aspect classification and polarity classification components are based on multiple filters CNNs as in (Ruder et al., 2016). OTE component is based on Bidirectional GRU architecture (similar to (Jebbara and Cimiano, 2016)). All the components are implemented with the keras (Chollet et al., 2015) library.

Since the size of the training data is relatively small, we attempt to enrich an input with prior knowledge to help the system to generalize better. In order to do so, we enrich word representation with semantic lexicon features[5], which are encoded as one-hot vector of dimension 100 and concatenated with word embedding. These new word representations are fed to the same pipeline as *baseline-2*. We'll refer to this system as *baseline-2'*.

Both *baseline-2* and *baseline-2'* are initialised with pre-trained word embeddings.

5.3 Baseline Results

Common ressources between all baselines are pre-trained word embeddings and semantic lexicon. We use word2vec (Mikolov et al., 2013) 300-dimensional Google News word embeddings, on which some"noise" filtering has been performed. Semantic lexicon was created semi-automatically using existing polarity lexicons and capitalizing on the annotated vocabulary present in the SemEval

[4]https://taku910.github.io/crfpp/

[5]This is close to the idea of *sentic features* (Jebbara and Cimiano, 2016), integrating aspect categories and polarities, rather than *sentics*.

Model	Foursquare				
	s2	s1	s3	s1,3	s1,2,3
baseline-1	68.9	63.8	88.7	56.9	33.6
baseline-2	47.9	62.9	86.0	52.5	9.1
baseline-2'	47.7	62.7	86.1	52.6	8.8
	Semeval				
baseline-1	75.3	70.4	87.3	63.0	37.1
baseline-2	61.1	69.9	80.2	54.9	12.0
baseline-2'	61.0	68.8	78.7	53.8	11.8

Table 2: Performance of various baseline systems. s1: Aspect Category detection (F1), s2: Opinion Target Expression (F1), s3: Sentiment Polarity (Accuracy). s1,3: Aspect,Polarity (F1), s1,2,3: Aspect,OTE,Polarity (F1).

ABSA datasets. It contains ∼1000 words with aspect categories and/or polarities associated to each word.

Results for all the baselines are summarized in the table 2. Note, that for *baseline-2,2'*, we report an average performance after executing the whole pipeline 10 times.

First, we observe an important performance drop in aspect prediction (tasks s2, s1) for the new Foursquare dataset for both baselines. This is of course related to the fact that this dataset is different from the one the training has been performed on. Thus, the aspects may not be expressed in the same way, style of the reviews are different[6]. However, for polarity prediction we observe better results on Foursquare dataset than on Semeval dataset. It can be explained by shorter length of Foursquare comments, resulting in less aspect mentions per sentence (rarely more that one opinionated term per sentence), and thus less ambiguity in polarity prediction.

The second observation is a pretty low overall pipeline performance (s1,3 and s1,2,3). Although our *baseline-1* has pretty good performances on each individual task (best, or close to best official SemEval2016 results) when putting all together, it results in 63.0 F1-score on aspect-polarity tuples. The performance on <OTE, Aspect, Polarity> tuples drops down to F1 of 37.1. This evaluation procedure allows us to get an idea on what would be "real-world" system performance, and also indicates the capacities and limitations of the system.

[6]a lot of emojis are used in Foursquare dataset, but not in Semeval dataset

Finally, we note that *baseline-1* ("ressource rich" baseline) has the best performances from all the baselines we explored (as expected). The performances of *baseline-2* and *baseline-2'* are pretty close on the Semeval dataset, but *baseline-2* seems to perform slightly better.

6 Exploring Additional Ressources for Adaptation

One of the natural resources to explore for system adaptation is a set non-annotated reviews. In our case, we exploit all Foursquare reviews in English we have access to.

6.1 Domain Specific Embeddings

First, we learn domain dependent words embeddings (300-dimensional) on the Foursquare restaurant data using Gensim (Řehůřek and Sojka, 2010) implementation of word2vec. We filtered out the words occurring less than 5 times, and used a context window of 10 words, which resulted in 60K word embeddings.

6.2 Weakly Supervised Lexical Acquisition

Among other components, our system relies on semantic lexical ressources encoding domain aspect and polarity vocabulary, that were developed semi-automatically, based on SemEval2016 training datasets. In order to enrich these lexicons, we have adapted a semantic clustering method described in (Pelevina et al., 2016)[7]. The core idea of this approach is to induce a sense inventory from existing word embeddings via clustering of ego-networks of related words. An ego network consists of a single node (ego) together with the nodes they are connected to and the edges between the connected nodes. Words referring to the same sense tend to have a large number of connections, and to be clustered together. The clustering is done with the Chinese Whispers algorithm (Biemann, 2006).

In the case of the present experiments, we initialize the algorithm with a set of seed words together with their semantic aspect (e.g *cider:drink*, *tikka:food*), in order to obtain clusters of aspect words. We used 60 seed words randomly selected from our existing semantic lexicon and learned clusters from Foursquare embeddings. Table 4

[7]This method was initially experimented for word sense disambiguation, but we directly adapted it for domain aspect lexicon creation

Model	Foursquare					Semeval				
	s2	s1	s3	s1,3	s1,2,3	s2	s1	s3	s1,3	s1,2,3
	baseline-1									
baseline-1	68.9	63.8	88.7	56.9	33.6	75.3	70.4	**87.3**	63.0	37.1
f_lex	**69.2**	64.1	**88.8**	57.1	33.8	**76.4**	70.4	86.6	**63.5**	**38.1**
f_emb	66.7	63.8	88.7	**57.3**	**34.1**	75.3	70.5	87.1	63.4	37.4
f_lex + f_emb	67.1	**64.3**	**88.8**	**57.3**	33.9	75.8	**70.7**	86.6	**63.5**	37.7
	baseline-2									
baseline-2	47.9	62.9	86.0	52.5	9.1	61.1	69.9	80.2	54.9	**12.0**
f_emb	**54.5**	**66.4**	**87.1**	**56.7**	**9.1**	**61.7**	**69.7**	**80.6**	**54.7**	11.3
	baseline-2'									
baseline-2'	47.7	62.7	86.1	52.5	8.8	61.1	68.8	78.7	53.8	**11.8**
f_lex	47.7	62.4	86.1	52.6	8.7	61.0	68.9	78.7	53.9	11.4
f_emb	**53.8**	**65.9**	**86.7**	**56.2**	**9.2**	**62.4**	**70.0**	**80.5**	**55.8**	11.4
f_lex + f_emb	**53.8**	65.8	**86.7**	**56.2**	**9.2**	**62.4**	69.9	**80.5**	**55.8**	11.4

Table 3: Experimental results with foursquare embeddings and automatically acquired lexicon

Seed:Aspect	**Aspect Cluster**
kimchi:food	kimchee, bulgogi, galbi, bibimbap, jigae, chigae, ...
waiter:service	waitress, server, hostess, nikki, melissa, kyle, kelly, ...
expensive:price	over-priced, pricey, costly pricy, cheap, spendy, ...

Table 4: Clusters learnt on Foursquare embeddings

gives some cluster examples. It's interesting to observe that we obtain a cluster of first names, often used to mention a waiter in Foursquare data, with semantic class *service*.

We use these clusters of aspect words by concatenating them to the existing lexicon of the system.

6.3 Experimental Results

We've performed following series of experiments (summarized in table 3): 1. *f_lex*: foursquare lexicon extending existing lexicon (for systems using lexicons); 2. *f_emb*: all baselines with foursquare embeddings replacing generic embeddings (GoogleNews-based) 3. *f_lex +f_emb*: combination of the previous two. We observe light improvements for *baseline-1* which are especially due to lexicon enrichment experiments. We think that Foursquare embeddings didn't bring expected improvements for *baseline-1* (embeddings are used only for OTE/s2 task, which in it's turn impacts s1 task), mostly because these embeddings are much smaller and we lose some non domain-specific knowledge when they replace GoogleNews embeddings.

The impact of embedding is opposite for *baseline-2* experiments. Foursquare pretrained embeddings bring important gains on Foursquare dataset thus moving *baseline-2* system above *baseline-1* for s1 evaluation. It also improves (although less) system performance on Semeval dataset. Automatically acquired lexicon on *baseline-2* systems seems to be very low. We plan to explore other ways to integrate this knowledge into deep learning framework.

7 Conclusion

In this work, we release a new ABSA dataset, in order to better assess state-of-the-art systems robustness; we also evaluate a full ABSA chain of various systems, to reflect end-to-end performances. We show that even for the systems with good performances on individual ABSA subtasks, an overall aspect/polarity F1 score drops down to 63.0. Evaluation of various baselines on the new dataset have shown that standard ABSA systems may suffer a significant decrease in performance, especially for aspect detection. We've experimented with light adaptation methods integrating in-domain embeddings and automatically acquired lexicons, and showed their impact on different systems. Both the new Foursquare ABSA dataset and the evaluation script of the full pipeline are distributed with the paper.

References

Chris Biemann. 2006. Chinese whispers: An efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, TextGraphs-1, pages 73–80, Stroudsburg, PA, USA. Association for Computational Linguistics.

François Chollet et al. 2015. Keras. https://keras.io.

Miguel Ángel García Cumbreras, Julio Villena-Román, Eugenio Martínez Cámara, Manuel Carlos Díaz-Galiano, Maria Teresa Martín-Valdivia, and Luis Alfonso Ureña López. 2016. Overview of TASS 2016. In *Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN co-located with 32nd SEPLN Conference (SEPLN 2016), Salamanca, Spain, September 13th, 2016.*, pages 13–21.

G. Ganu, N. Elhadad, and A. Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In *Proceedings of the 12th International Workshop on the Web and Databases*, Providence, Rhode Island.

Hussam Hamdan, Patrice Bellot, and Frederic Bechet. 2015. Lsislif: Crf and logistic regression for opinion target extraction and sentiment polarity analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 753–758, Denver, Colorado. Association for Computational Linguistics.

Soufian Jebbara and Philipp Cimiano. 2016. Aspect-Based Sentiment Analysis Using a Two-Step Neural Network Architecture. In *Semantic Web Challenges. Third SemWebEval Challenge at ESWC 2016. Revised Selected Papers*, volume 641, pages 153–170. Springer.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 437–442, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Ayush Kumar, Sarah Kohail, Amit Kumar, Asif Ekbal, and Chris Biemann. 2016. Iit-tuda at semeval-2016 task 5: Beyond sentiment lexicon: Combining domain dependency and distributional semantics features for aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1129–1135, San Diego, California. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119.

Maria Pelevina, Nikolay Arefyev, Chris Biemann, and Alexander Panchenko. 2016. Making sense of word embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 174–183.

Maria Pontiki, Dimitrios 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, Evgeny 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 '16, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. 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 and Dublin City University.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

S. Ruder, P. Ghaffari, and J. G. Breslin. 2016. INSIGHT-1 at SemEval-2016 Task 5: Deep Learning for Multilingual Aspect-based Sentiment Analysis. *ArXiv e-prints*.

Pontus Stenetorp, Sampo Pyysalo, Goran Topi, Tomoko Ohta, and Sophia Ananiadou. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 102–107.

Zhiqiang Toh and Wenting Wang. 2014. Dlirec: Aspect term extraction and term polarity classification system. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 235–240, Dublin, Ireland. Association

for Computational Linguistics and Dublin City University.

Julio Villena-Román, Janine García-Morera, Miguel Ángel García Cumbreras, Eugenio Martínez-Cámara, Maria Teresa Martín-Valdivia, and Luis Alfonso Ureña López. 2015a. Overview of TASS 2015. In *Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN co-located with 31st SEPLN Conference (SEPLN 2015), Alicante, Spain, September 15, 2015.*, pages 13–21.

Julio Villena-Román, Eugenio Martínez-Cámara, Janine García-Morera, and Salud M. Jiménez Zafra. 2015b. TASS 2014 - the challenge of aspect-based sentiment analysis. *Procesamiento del Lenguaje Natural*, 54:61–68.

Joachim Wagner, Piyush Arora, Santiago Cortes, Utsab Barman, Dasha Bogdanova, Jennifer Foster, and Lamia Tounsi. 2014. Dcu: Aspect-based polarity classification for semeval task 4. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 392–397, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Michael Wojatzki, Eugen Ruppert, Sarah Holschneider, Torsten Zesch, and Chris Biemann. 2017. GermEval 2017: Shared Task on Aspect-based Sentiment in Social Media Customer Feedback. In *Proceedings of the GermEval 2017 Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, pages 1–12, Berlin, Germany.