Proceedings of CSR-NLP I @LREC 2022, pages 35–39
Marseille, 25 June 2022

MobASA: Corpus for Aspect-based Sentiment Analysis and Social Inclusion in the Mobility Domain

Aleksandra Gabryszak, Philippe Thomas
Deutsches Forschungszentrum für Künstliche Intelligenz
Alt-Moabit 91c, 10559 Berlin, Germany
{aleksandra.gabryszak, philippe.thomas}@dfki.de

Abstract

In this paper we show how aspect-based sentiment analysis might help public transport companies to improve their social responsibility for accessible travel. We present MobASA: a novel German-language corpus of tweets annotated with their relevance for public transportation, and with sentiment towards aspects related to barrier-free travel. We identified and labeled topics important for passengers limited in their mobility due to disability, age, or when travelling with young children. The data can be used to identify hurdles and improve travel planning for vulnerable passengers, as well as to monitor a perception of transportation businesses regarding the social inclusion of all passengers. The data is publicly available under: https://github.com/DFKI-NLP/sim3s-corpus

Keywords: Aspect-based Sentiment Analysis, Crowdsourcing, Mobility, Social Inclusion, Social Responsibility

1. Introduction

Social inclusion is of great importance for building stable societies and public transportation companies play a particularly substantial role for ensuring equal participation in society. Unfortunately, accessing trains, buses, or stations is often a challenge for people limited in their mobility due to a physical or cognitive impairment, age or when travelling with young children. It is important to enable those groups to use public transport in a self-reliant way by providing facilities (lifts or ramps for walking disabled people, etc.), services (visual info for deaf and acoustic info for blind people, etc.), as well as systems informing about the state of these forms of assistance (if lifts are available, etc.) in order to identify unexpected hurdles and improve travel planning. Natural language processing provides means to aid such systems by the automatic extraction of information from texts about the condition of relevant facilities and services. For example, given the input text:

> *A lift at the Berlin Hbf station has been already defect for two days! This is really annoying!*

it would be helpful to have a system that is able to determine that (a) the availability of lifts at public transportation stations is mentioned and (b) their state is described as malfunctioning. The extracted information can be used to inform transport operators as well as passengers limited in their mobility about a problem of a specific facility, trigger a process solving or mitigating the problem (e.g. fixing broken lifts, proposing an alternative traveling route).

In this paper we devote our attention to the question of supporting such systems by adapting the aspect-based sentiment analysis task. Sentiment analysis aims at extracting and quantifying subjective information. A standard version of the sentiment detection classifies the sentiment of a whole sentence, while the aspect-based sentiment analysis (ABSA) focuses on the sentiment towards predefined aspects such as specific products or services. Therefore, ABSA allows a more fine-grained mining of opinions. We cast our problem of extracting information on the state of facilities and services relevant to the barrier-free accessibility of public transport as an aspect-based sentiment task. We consider facilities and services as main aspects, their properties as aspect subcategories, and statements about those properties as phrases potentially expressing or implying a sentiment. For the example above, we assume a main category *lift*, a subcategory *availability* (of the lift), and a negative sentiment towards the aspect *Lift#Availability*.

As a result of our work we present MobASA, a German-language dataset for the detection of sentiment towards aspects relevant for users of public transport limited in their mobility. To the best of our knowledge there is no other dataset, English or German, which covers the topic of travel accessibility in a fine-grained way. Our contributions are:

- We provide a publicly available German-language dataset for the detection of aspect-based sentiment towards barrier-related aspects in the public transport domain.
- The dataset can benefit building inclusive public transportation systems as described in the introduction. Therefore, we add to research aiming to deploy various NLP tasks in support of equality and social responsibility of businesses.

2. Related Work

Aspect-based Sentiment The annotated datasets for developing ABSA models are still scarce, and they mostly cover only the standard domain of product or service reviews (e.g. SCARE corpus by (Sänger et al., 2016), SemEval 2015 by (Pontiki et al., 2015), USAGE by (Klinger and Cimiano, 2014), GESTALT by (Ruppenhofer et al., 2014)). In contrast, the GermEval 2017 dataset (Wojatzki et al., 2017) comprises social media texts annotated with opinions on the biggest railway company in Germany. It lists barrier-free accessibility as one coarse-grain aspect, however more refined labels are needed to model information needs of different target groups (e.g. blind vs. deaf people).

Recent neural approaches based on pre-trained language models (e.g., BERT (Devlin et al., 2019)) have shown impressive results for the task when fine-tuned on supervised datasets. However, the state-of-the-art transformer based ABSA models currently achieve an F1-score of only 0.53 on the GermEval 2017 dataset (Aßenmacher et al., 2021) and 0.61 on the SemEval 2016 laptop-dataset (Pontiki et al., 2016; Li et al., 2019), for example, meaning there is still much room for improvement.

Inclusive NLP Natural language processing (NLP) technologies already support the efforts of inclusion in various domains, for example, sign-language-to-text translation systems (Nunnari et al., 2021) to benefit deaf people, domain-specific translation systems to support migrants when communicating with authorities (Xu et al., 2018), as well as applications predicting readability to help content providers with adjusting their published texts to the needs of people with cognitive disabilities (Evans et al., 2016). The systems target mostly language and communication barriers.

3. Dataset

3.1. Data Collection

To collect the relevant data we crawled German-language tweets based on a predefined list of 11 Twitter channels of public transportation companies, channels related to barrier-free accessibility as well as a set of 68 keywords relevant to the barrier-free travel of handicapped passengers, older people or parents with small children. The list contained German-language keywords equivalent to words such as: *barrier-free, escalator, guiding system for the blind*, etc. We collected 3,128,639 tweets between 2019-2021, and from that data we sampled tweets for the annotation.

3.2. Annotation Schema and Guidelines

The MobASA labels structure and annotation guidelines are partly based on the instructions of GermEval 2017 and SemEval 2015 datasets. The set of meta-labels (relevance, sentiment, category, polarity, from, to) as well as the XML data format originate from GermEval 2017.

Relevance for Public Transportation Each tweet has binary labelling regarding its `relevance` to public transport. The relevance value is *true*, if a tweet contains any phrase related to public passenger transport of any type. For example, the text in Figure 1 contains mentions of a metro station.

Aspect-based Sentiment We defined a base catalog of 19 aspect `categories` relevant to the barrier-free travel. We included aspects important for the walking disabled passengers, people with a vision or hearing impairment, as well as the elderly, and parents traveling with small children. The category catalog is based on interviews with those target groups, guidelines for travel accessibility by the government and interest groups, information provided by the biggest German railway operator, as well as topics mentioned in our data. Each aspect category consists of two parts: a main aspect and its subcategory. The main aspect references mostly a specific assistance form (facility or service) such as lift, lighting or acoustic info. The subcategory captures various relevant features of the main aspect such as its availability among others. The subcategory might also be labelled *Main* if no multiple, specific subcategories are identified. We defined up to two subcategories for each aspect. The category *Others* was annotated if an unanticipated or less frequent but relevant topic was not covered in the base catalog. For example, very short-term announcements of platform changes for departing trains might result in people impaired in their mobility missing their train. Examples of the annotation of various subcategories for tweets, which referenced a main aspect are given in Table 2.

Furthermore, each aspect is annotated along with a `polarity` value *neutral*, *positive* or *negative*. The value indicates either stated sentiment towards an aspect or, more broadly, it indicates the described state of that aspect. For example, the polarity of the category *GeneralBarrier#Main* is usually positive if a station or a train is stated as being handicapped accessible, negative if it is not, and neutral if the degree of accessibility is stated as unknown or is described as neither positive nor negative for other reasons. The texts might contain opinionated statements such as *bad* or *good*, however, this is not required, i.e we also accept polarities implied by the state of the aspect (e.g. a faulty lift implies a negative polarity as in the example in Figure 1).

Furthermore, we asked to annotate the value regarding the most recent described or announced state of the aspect, i.e. if the lighting was faulty, but it is stated as already repaired, then the sentiment is positive. This approach was chosen with the aim in mind to support systems which focus on solving the latest problems when using public transport.

The `target` of an annotated aspect and its polarity is a text span referencing the main aspect, e.g. phrases *Aufzug* denoting the main aspect *Lift*. The offsets of the target span are marked by the labels `from` and `to`.

Document-level Sentiment Each tweet is labeled with a document-level `sentiment`. Its value aggregates the polarities of the opinions in a given text. If the polarity set is {*positive*, *neutral*} or {*negative*, *neutral*} then the document-level sentiment is set to positive or negative, respectively, otherwise the value is *neutral* (as illustrated by the example in Figure 1). If a text is irrelevant, then the document-level sentiment is *neutral* by default.[1]

3.3. Annotation process and quality

Expert annotation A subset of tweets is fully annotated by trained experts using the platform *Inception*[2]. The final expert subset of the corpus includes only annotations, for which two annotators agreed or the disagreement was resolved by the third annotator. The annotation is based on guidelines, which were developed in an iterative process and take into account discussions with the experts. The annotators were given definitions of relevance and the aspects along with multiple examples. The annotation of aspect-

[1] We follow in our approach the heuristic used in GermEval2017 data.

[2] https://inception-project.github.io/

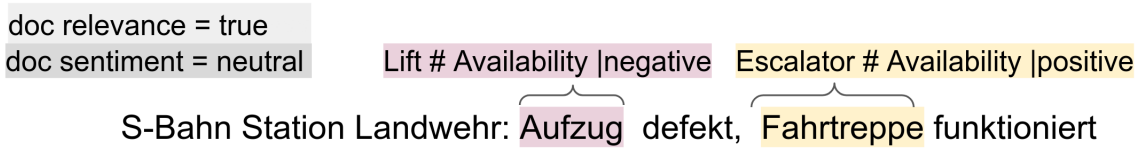

Figure 1: Example of a German-language tweet relevant for public transportation and containing negative sentiment towards the aspect *Lift#Availability* and positive sentiment towards the aspect *Escalator#Availability*. (Text in EN: *S-Bahn Station Landwehr: Lift defect, Escalator is working.*)

aspect	description
AccidentsMobilityGroups#Main	risks of injury for people with limited mobility (e.g. falls of wheelchair users into track bed)
AcousticSignal#Main	acoustic signals for blind people (acoustic warning or signals for finding train doors, etc.)
ConstructionSite#Main	construction sites and their impact on the public transport (e.g. accessing of stations)
Demonstration#Main	demonstrations and their impact on the public transport (e.g. accessing of stations)
Escalator#Availability	operational status of escalators (e.g. if they exist and function properly)
Escalator#Tidiness	cleanliness of escalators (also smell or similar)
GeneralBarrier#Main	general mentions of barrier-free accessibility in public transport
GroundLevelAccess#Main	ground level access to stations or vehicles of public transport
InfoAcoustic#Availability	availability of announcements or operational state of loudspeakers
InfoDisplay#Availability	availability of displayed information or operational state of display boards
Info#Others	availability and quality of information on public transport in apps, e-mails, etc.
Lift#Availability	operational status of lifts
Lift#Tidiness	cleanliness of lifts (also smell or similar)
Lighting#Availability	operational status of lighting
Ramp#Availability	operational status of ramps
Security#Main	security at stations (e.g. important for older or handicapped people)
SpaceMobilityGroups#Main	space available for people limited in their mobility (e.g. wheelchair bay)
TactileContrastOrientation#Main	tactile or high-contrast guiding routes for blind people, info in braille, etc.
main category#Others	not anticipated or less frequent subcategories (e.g. InfoDisplay#Others for correctness of displayed info)
BarrierOthers#Main	other topics related to barrier-free accessibility (e.g. assistance during traveling)

Table 1: Definition of aspect categories related to barrier-free accessibility

based sentiment was only considered for the data annotated as relevant. The inner-annotator agreement for the various annotation layers is: 1) relevance: Cohen's $\kappa = 0.96$, 2) aspect-based sentiment: Cohen's $\kappa = 0.73$ on annotated tokens only. Therefore we achieved nearly perfect agreement in the relevance annotation and substantial agreement in the aspect-based sentiment annotation.

Crowdsourcing An additional subset of tweets was annotated by crowdworkers using the platform *Crowdee*[3]. First, the workers labelled tweets as relevant or irrelevant for the public transportation topic. Subsequently, the tweets were annotated regarding aspect-based sentiment. In order to choose relevant candidates for the aspect annotation, first we sampled tweets already labeled or automatically determined as relevant for public transportation. For the automatic detection we systematically collected phrases referring to transportation types from a subset of relevant tweets, and used those phrases to filter the potentially relevant data. In the next step we automatically pre-annotated text spans with the main aspect category (e.g. word *Fahrstuhl* with main category `Lift`) by matching text spans to target strings annotated in the expert subset. Then we showed crowdworkers texts, where a pre-annotated main aspects were highlighted, and we asked if a specific subcategory regarding the highlighted aspect is discussed in a given text, and if so with which polarity. The task was designed as a multiple-choice questionnaire. For the crowdsourcing we focused on aspects most relevant to various target groups (e.g. *Escalator#Availability*), and excluded rare or less relevant aspects (e.g. *TactileContrastOrientation#Main*, *Lift#Tidiness*). For the annotation of both tasks we provided short guidelines as well as many examples. Each tweet was processed by two workers. To ensure a higher quality of the crowdsourcing process we prepared a qualification test, inserted trapping questions, and set a minimum time for solving the task. We blocked all users, who failed the tests from further tasks. Finally we included only annotations, for which two workers agreed on. We reviewed a sample of the crowd-sourced labels included in the final data. We estimated the accuracy of the relevance labels as very high having 99.6% correct labels of 1000 sampled tweets. Regarding aspect-based sentiment we reviewed 1950 answers and estimate the accuracy as high based on the 84.9% correct labels.

3.4. Data Statistics

We provide dev and test split as well as two versions of the train set (Table 3). The dev and test and $\text{train}_{\text{BASE}}$ split contain data, in which all relevant tweets are annotated by the experts, and the irrelevant data is partially labeled by the crowdworkers. We also publish an extended version of the train corpus, $\text{train}_{\text{PLUS}}$, which additionally contains the

[3]https://www.crowdee.com/

aspect	polarity	example
InfoDisplay#Availability	neutral	*Sag mir mal, wenn es geht, ob die Anzeigetafeln am Hbf wieder gehen!* :D (*Tell me if the displayboards at the main station are working again! :D)*
InfoDisplay#Availability	positive	*Die Anzeigetafeln am Hauptbahnhof Bremen laufen wieder* (*The displayboards at Bremen Central Station are functioning again)*
InfoDisplay#Availability	negative	*S2 08:02 ab Bernau fährt nicht weil? [...] keine Anzeige.* Scheiße! (*S2 08:02 from Bernau is not coming because? [...] no info displayed. Crap!*)
InfoDisplay#Others	negative	*Sbahn fällt 3x aus, [...] schrift auf anzeigetafle ist verkehrt herum* (*Sbahn canceled 3x time, [...] text on displayboard is upside down*)
(no relevant aspect)	-	*@jbnieo Komm mit der Bahn so um 12.06 an, lass uns dann bei der Anzeigetafel treffen.* (*@jbnieo Arrive by train around 12.06, then let's meet at the displayboard.*)

Table 2: Examples of the annotation of texts containing the main aspect *InfoDisplay* (original texts and English translations)

crowd-sourced annotations of the aspect-based sentiment.

The inclusion of primarily expert annotation in dev and test set ensures a more robust selection and evaluation of the models, since the expert annotation introduces less noisy labels. That approach follows the suggestions of a careful design of the test data to not misrepresent model performance (Alt et al., 2020; Bowman and Dahl, 2021).

	total	dev	test	$\text{train}_{\text{BASE}}$	$\text{train}_{\text{PLUS}}$
docs	29446	4176	4192	12510	21078
relevant	18378	1960	1899	5951	14519
aspects	13533	1205	1150	3572	11178

Table 3: Statistics of the data splits

Relevance The binary relevance labels are relatively equally distributed in dev, test and $\text{train}_{\text{BASE}}$ (Table 3). In $\text{train}_{\text{PLUS}}$ the relevant docs are dominant, since we selected only relevant classes for the crowdsourcing of aspect annotation. Therefore, the second split should primarily be used for the detection of aspect-based sentiment.

Aspect and Sentiment Table 4 shows 10 most frequent aspects. For some main concepts such as *TactileContrastOrientation* for blind people we almost found no data. In contrast, other main aspects such as *InfoAcoustic* or *Escalator* are often mentioned in text, however not in a context relevant to barrier-free accessibility. Also subcategories such as *Tidiness* of lifts or escalators are rarely mentioned.

Table 5 shows the distribution of document-level and aspect-level sentiment values. The neutral values build the majority of classes on the document-level, however it is due to the annotation of irrelevant tweets with neutral sentiment as default. For the relevant data negative sentiment is the most dominant on document- and span-level. We also observed, that mostly the expressed sentiment refers to the current or past state of the aspect. However for the category *GeneralBarrier#Main* the positive sentiment often refers to the future state, e.g. the future barrier-free accessibility of stations is announced.

aspect	total	expert	crowd
GeneralBarrier#Main	3010	807	2203
Lift#Availability	2918	1586	1332
Escalator#Availability	1615	769	846
InfoDisplay#Availability	1545	351	1194
ConstructionSite#Main	1069	103	966
Lighting#Availability	786	509	277
InfoAcoustic#Availability	686	231	455
Ramp#Availability	434	101	333
InfoDisplay#Others	349	349	0
Demonstration#Main	232	232	0
others	889	889	0
total	13533	5927	7606

Table 4: Statistics of 10 most frequent aspect categories

polarity	doc level	span level
neutral	19660	1652
positive	1968	2361
negative	7818	9520
total	29446	13533

Table 5: Statistics of the aspect-based sentiment

4. Conclusion and Future Work

Most of the inclusive NLP systems focus on overcoming communication barriers. In contrast, we show how NLP can be used by public transportation businesses to mitigate barriers resulting from broken travel facilities or services, and in result to support inclusion of all passengers. We presented a corpus of tweets annotated with sentiment towards aspects related to barrier-free travel. In future work, we want to refine the aspect catalog, and integrate the detection of aspect location and time, to which the sentiment refers.

5. Acknowledgements

This research was partially supported by the Federal Ministry of Transport and Digital Infrastructure (BMVI) through the project SIM3S (19F2058B) and by the German Federal Ministry of Education and Research (BMBF) through the project BIFOLD (01IS18025E).

6. Bibliographical References

Alt, C., Gabryszak, A., and Hennig, L. (2020). TACRED revisited: A thorough evaluation of the TACRED relation extraction task. *CoRR*, abs/2004.14855.

Aßenmacher, M., Corvonato, A., and Heumann, C. (2021). Re-evaluating germeval17 using german pre-trained language models. In *Proceedings of the Swiss Text Analytics Conference 2021*.

Bowman, S. R. and Dahl, G. E. (2021). What will it take to fix benchmarking in natural language understanding? *CoRR*, abs/2104.02145.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Evans, R., Yaneva, V., and Temnikova, I. (2016). Predicting reading difficulty for readers with autism spectrum disorder.

Klinger, R. and Cimiano, P. (2014). The USAGE review corpus for fine grained multi lingual opinion analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2211–2218, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Li, X., Bing, L., Zhang, W., and Lam, W. (2019). 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, November. Association for Computational Linguistics.

Nunnari, F., España-Bonet, C., and Avramidis, E. (2021). A data augmentation approach for sign-language-to-text translation in-the-wild. In *Proceedings of the 3rd Conference on Language, Data and Knowledge. Conference on Language, Data and Knowledge (LDK-2021), September 1-4, Zaragoza/Hybrid, Spain*, volume 93 of *OpenAccess Series in Informatics (OASIcs)*. Dagstuhl publishing, 9.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., and Androutsopoulos, I. (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, June. Association for Computational Linguistics.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S. M., and Eryiğit, G. (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, June. Association for Computational Linguistics.

Ruppenhofer, J., Klinger, R., Struš, J. M., Sonntag, J., and Wiegand, M. (2014). Iggsa shared tasks on german sentiment analysis (gestalt).

Sänger, M., Leser, U., Kemmerer, S., Adolphs, P., and Klinger, R. (2016). SCARE — the sentiment corpus of app reviews with fine-grained annotations in German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1114–1121, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Wojatzki, M., Ruppert, E., Holschneider, S., Zesch, T., and Biemann, C. (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.

Xu, F., Uszkoreit, H., Schmeier, S., and Ayach, A. (2018). Fahum heißt verstehen: Eine flüchtlings-app für soforthilfe und integration. In Aljoscha Burchardt et al., editors, *IT für soziale Inklusion: Digitalisierung – Künstliche Intelligenz – Zukunft für alle*, pages 151–154. De Gruyter Oldenbourg.