GermEval 2017: Shared Task on Aspect-based Sentiment in Social Media Customer Feedback

Michael Wojatzki[†] and **Eugen Ruppert**[‡] and **Sarah Holschneider**[◇] and **Torsten Zesch**[†] and **Chris Biemann**[‡]

[†]**Language Technology Lab**
CompSci and Appl. CogSci
Universität Duisburg-Essen
http://www.ltl.uni-due.de/

[◇]**Technische Universität Darmstadt**
http://www.tu-darmstadt.de

[‡]**Language Technology Group**
Computer Science Department
Universität Hamburg
http://lt.informatik.uni-hamburg.de

Abstract

This paper describes the GermEval 2017 shared task on Aspect-Based Sentiment Analysis that consists of four subtasks: relevance, document-level sentiment polarity, aspect-level polarity ad opinion target extraction. System performance is measured on two evaluation sets – one from the same time period as the training and development set, and a second one, which contains data from a later time period. We describe the subtasks and the data in detail and provide the shared task results. Overall, the shared task attracted over 50 system runs from 8 teams.

1 Introduction

In a connected, modern world, customer feedback is a valuable source for insights on the quality of products or services. This feedback allows other customers to benefit from the experiences of others and enables businesses to react on requests, complaints or recommendations. However, the more people use a product or service, the more feedback is generated, which results in the major challenge of analyzing huge amounts of feedback in an efficient, but still meaningful way.

Recently, shared tasks on Sentiment Analysis have been organized regularly, the most popular are the shared tasks in the SemEval framework (Pontiki et al., 2015; Pontiki et al., 2016). And even though the number of domains and languages is growing with each iteration, there has not existed a large public sentiment analysis dataset for German until now.

To fill this gap, we conducted a shared task[1] on automatically analyzing customer reviews and news about "Deutsche Bahn" – the major German public train operator, with about two billion passengers each year. This is the first shared task on German sentiment analysis that provides a large annotated dataset for training and evaluating machine learning approaches. Furthermore, it features one of the largest datasets for sentiment analysis overall, containing annotations on almost 28,000 short documents, more than 10 times of the training instances in the largest set to date (from SemEval-2016, task 5 'Arabic Hotels').

2 Task Description

The shared task features four subtasks, which can be tackled individually. They are aimed at realizing a full classification pipeline when dealing with web data from various heterogeneous sources. First, in Subtask A, the goal is to determine whether a review is relevant to our topic. In real life scenarios this task is necessary to filter irrelevant documents that are a by-catch of the method of collecting the data. Second, Subtask B is about inferring a customer's overall evaluation of the Deutsche Bahn based on the given document. Here, we support a use-case in which e.g. a manager is interested how well or badly the offered services are perceived overall. Third, Subtask C addresses a more fine-grained level and aims at finding the particular kind of service, called aspect, which is referred to positively or negatively. Finally, in Subtask D the task is to identify the actual expressions that verbalize the evaluations covered in Subtask C, commonly known as opinion target expression (OTE) identification.

2.1 Subtask A: Relevance Classification

The first subtask is used to filter incoming documents so that only the relevant and interesting ones are processed further. The term *Bahn* can refer to many different things in German: the rails, the train, any track or anything that can be laid in straight

[1] Documents and description of the GermEval 2017 shared task are available on the task website: https://sites.google.com/view/germeval2017-absa/

Proceedings of the GermEval 2017 – Shared Task on Aspect-based Sentiment in Social Media Customer Feedback, pages 1–12,
Berlin, Germany, September 12, 2017.

lines. Therefore, it is important to remove documents about e.g. the *Autobahn* (highway). This is similar for other query terms that are used to monitor web sites and microblogging services.

In Subtask A, the documents have to be labeled in a binary classification task as relevant (true) or irrelevant (false) for Deutsche Bahn. Below is a relevant document about bad behavior in a train, and an irrelevant document about stock exchange developments.

true Ehrlich die männer in Der *Bahn* haben keine manieren? (Seriously, the men in those trains have no manners!)

false Aus der Presseschau: Japanische S-Bahn wird mit Spiegelwaggons 'unsichtbar '(Review: Japanese urban railway becomes 'invisible' thanks to reflecting wagons)

2.2 Subtask B: Document-level Polarity

In Subtask B, systems have to identify, whether the customer evaluates the service of the railway company, be it e.g. travel experience, timetables or customer communication as positive, negative, or neutral. During data acquisition, annotators provided more complex aspect-level annotations as used in Subtasks C and D. Document level sentiment polarity in Subtask B is computed from the individual aspect polarities in the document: If there is a mixture between neutral and positive/negative, the documents are classified as positive/negative. If there are two opposing polarities (positive and negative), the overall sentiment is set to neutral.

2.3 Subtask C: Aspect-level Polarity

For Subtask C, participants are asked to identify all aspects in the document. Each aspect should be labeled with the appropriate polarity label. Since, in the annotations, it was possible to label multiple tokens with the same aspect, multiple mentions of the same aspect are possible. The example below shows a mixed sentiment in a document that is presented as a dialogue.

The positive aspect is the end of a strike – *Streik beendet*. The negative aspect in this document are the tickets, which are getting more expensive – *die Tickets teurer*. Thus, in the given example, the task is to identify the aspects (and their polarity) in the following way: Ticketkauf#Haupt:negative, Allgemein#Haupt:positive.

	Sentiment	Example
German	negative	Re: Ingo Lenßen Guten morgen Ingo...bei mir kein regen aber bahn fehr wieder nicht....liebe grusse ....
	positive	Re: DB Bahn Danke, hat sich gerade erledigt. Das Team hat mich per E-Mail kontaktiert. Danke trotzdem für die prompte Antwort:-)
	neutral	Kann man beim DB Navigator (APP) auch Jugend/Kinder Karten buchen?
English	negative	Re: Ingo Lenßen Good morning Ingo...No rain where I am but no trains again. Best wishes ....
	positive	Re: DB Bahn Thanks, sorted. I was contacted by the team. Anyways, thanks for replying so fast :-)
	neutral	Can you book concessions/child tickets using the DB Navigator (App)?

Table 1: Example for Document Sentiment

	Sentiment	Example
German	positive	Alle so 'Yeah, Streik beendet'
	negative	Bahn so 'Okay, dafür werden dann natürlich die Tickets teurer 'Alle so 'Können wir wieder Streik haben?'
English	positive	Everybody's like 'Yeah, strike's over'
	negative	Bahn goes 'Okay, but therefore we're going to raise the prices 'Everybody's like 'Can we have the strike back?'

Table 2: Example for Document Sentiment

The aspect classification was provided by the data analysis from Deutsche Bahn and was refined during the annotation process.

2.4 Subtask D: Opinion Target Extraction

For the last subtask, participants should identify the linguistic expressions that are used to express the aspect-based sentiment (Subtask C). The opinion target expression is defined by its starting and ending offsets. For human readability, the target terms are also present in the data as well.

An example is given in Listing 1. In this document, the task is to identify the target expression *fährt nicht* (does not drive/go), which is an indication of an irregularity in the operating schedule.

While the data set is available in both TSV and XML formats (see Section 3.4), Subtask D can only be done using the XML format, as the spans of the opinion target expression are not available in the

```
<Document>
  <text>@m_wabersich IC 2151? Der fährt nicht. Ich habe Ihnen die Alternative
     bereits genannt. /je</text>
  <Opinions>
    <Opinion aspect="Sonstige_Unregelmässigkeiten#Haupt" from="26" to="37" polarity
       ="negative" target="fährt nicht"/>
   </Opinions>
</Document>
```

Listing 1: Example document for Subtask D. Translation: @m wabersich IC 2151? It does not run. I already have told you about an alternative. aspect=miscellaneous irregularities#Main, target "does not run".

document-based TSV format. For more detail, see the next section.

3 Dataset

3.1 Data Collection

The data was crawled from the Internet on a daily basis with a list of query terms. We filtered for German documents and focused on social media, microblogs, news, and Q&A sites. Besides the document text, meta information like URL, date, and language was collected as well.

In the project context, we received more than 2.5 million documents overall, spanning a whole year (May 2015–June 2016), so that we could capture all possible seasonal problems (holidays, heat, cold, strikes) as well as daily problems such as delays, canceled trains, or unclear information at the train stations. From this large amount of documents, we sampled from each month approximately 1,500 documents for annotation. Since the word-list-based relevance filtering is very coarse and a lot of irrelevant documents were present in the initial samples, e.g. questions about the orbit of the moon (*Mondumlauf**bahn**, lunar orbit*) or mentions of air draft (***zug**ig, drafty*), we trained a baseline SVM classifier to perform pre-filtering and increase the number of relevant and interesting documents per split. The annotated data is used for the training, development, as well as for a **synchronic test set**.

Additionally, to test the robustness of the participating systems, we created a **diachronic test set**, which was (pre-)processed and annotated in the same manner, using data from November 2016 to January 2017.

3.2 Annotation

For annotation, we used WebAnno (de Castilho et al., 2016). Annotators were asked to perform the full annotation of every document assigned to them. To keep the individual tasks manageable, we split the annotation tasks into chunks of 100 short documents, which could be completed in 1–2 hours by an annotator.

The annotation task consisted of first labeling the document relevance. For relevant documents, the annotators had to identify the aspect targets (spans of single or multiple tokens) and label them with one of 19 aspects and, if identifiable, with one of overall 43 sub-aspects.

Relevant documents that did not contain a clear aspect target expression could also be assigned a document-level aspect annotation. The polarity words for each aspect target were annotated as well. If they were not part of the OTE (as e.g. *Verspätung – delay*, which is inherently negative), they were connected with the aspect-bearing word using a relation arc. These annotations have not been distributed as part of this challenge, but will be made available afterwards. Expressions of the same aspect were also connected from left to right.

The annotation team consisted of six trained student annotators and a supervisor/curator. Every document was annotated by two annotators in differing pairings. The curator checked the documents for diverging annotations and decided on the correct one using WebAnno's curation interface. Furthermore, she was also able to add new annotations, in case the others missed some expressions. In weekly feedback sessions, the team talked about new problems and added the results to the annotation guidelines.[2] This led to consistent improvements of inter-annotator agreement over time, see Table 3 and Figure 1. The overall lower agreements for the Relevance classifications are due to the difficulty of deciding between irrelevant documents and documents without explicit sentiments.

[2]The German annotation guidelines are available at: http://ltdata1.informatik.uni-hamburg.de/germeval2017/Guidelines_DB_v4.pdf

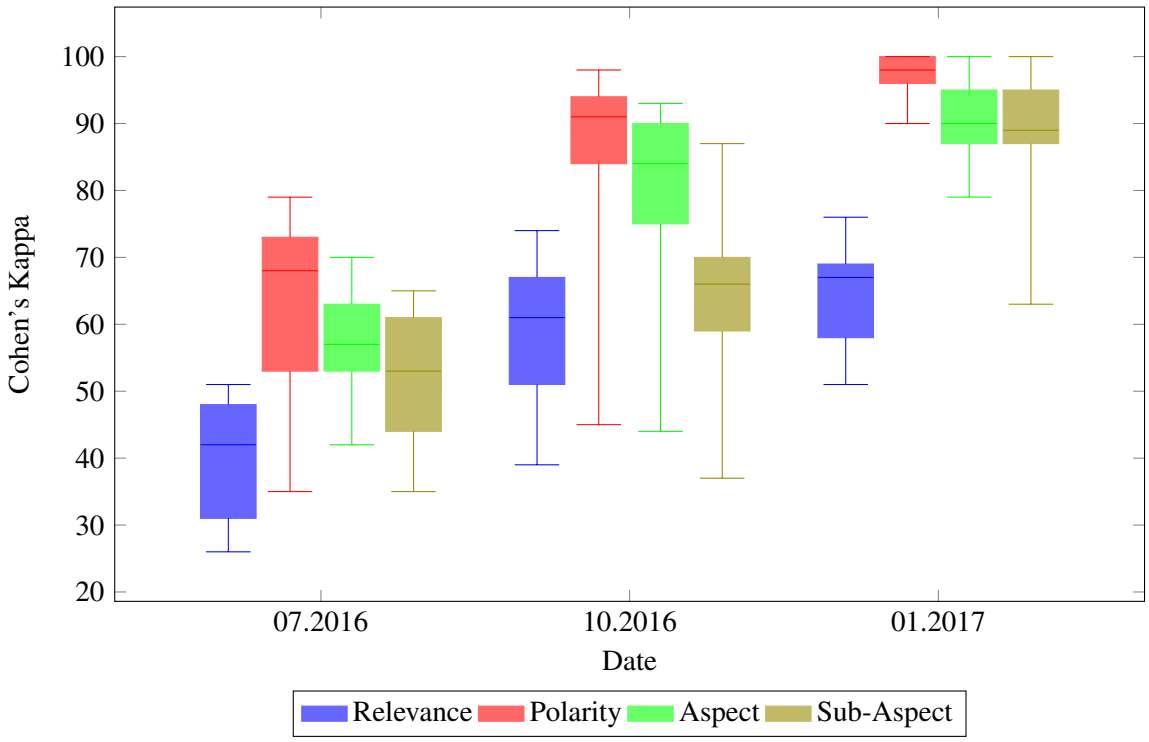

Figure 1: Development of the inter-annotator agreement over time

Date	07.2016	10.2016	01.2017
Relevance	0.26–0.51	0.39–0.74	0.51–0.76
Polarity	0.35–0.79	0.45–0.97	0.90–1.00
Aspect	0.42–0.70	0.44–0.93	0.79–1.00
Sub-Aspect	0.35–0.65	0.37–0.87	0.63–1.00

Table 3: Development of the inter-annotator agreement ranges (Cohen's kappa)

3.3 Splits

We obtained about 26,000 annotated documents for the main dataset and about 1,800 documents for the diachronic dataset. We split the main dataset into a training, development and test set using a random 80%/10%/10% split. The number of documents for each split is shown in Table 4. The dataset can be downloaded from: `http://ltdata1.informatik.uni-hamburg.de/germeval2017/`

Tables 5–7 show the label distributions for the subtasks. There is always a clear majority class, which leads to strong baselines. This is especially apparent for Subtask C (Table 7), where the most frequent label *Allgemein (General)* is almost 10 times as frequent as the second frequent label.

train	dev	test_syn	test_dia
19,432	2,369	2,566	1,842

Table 4: Number of documents in data splits

Dataset	true	false
Training	16,201	3,231
Development	1,931	438
Test_syn	2,095	471
Test_dia	1,547	295

Table 5: Relevance Distribution in Subtask A data

Dataset	negative	neutral	positive
Training	5,045	13,208	1,179
Development	589	1,632	148
Test_syn	780	1,681	105
Test_dia	497	1,237	108

Table 6: Sentiment Distribution in Subtask B data

3.4 Formats

We utilize an XML format that is similar to the format used in SemEval-2016 task on ABSA (Task 5) (Pontiki et al., 2016). Each `Document` element contains the original URL as the document id and

	Training	Development	Test_syn	Test_dia
	11,191 Allgemein	1,363 Allgemein	1,351 Allgemein	1,008 Allgemein
	1,240 Zugfahrt	140 Zugfahrt	178 Sonstige...	144 Zugfahrt
	1,007 Sonstige_Unregelmässigkeiten	108 Atmosphäre	160 Zugfahrt	138 Sonstige...
	819 Atmosphäre	102 Sonstige...	112 Atmosphäre	72 Connectivity
Top 10	417 Ticketkauf	51 Ticketkauf	75 Ticketkauf	42 Atmosphäre
Aspects	296 Service_und_Kundenbetreuung	37 Sicherheit	51 Sicherheit	29 Ticketkauf
	278 Sicherheit	29 Service...	42 Service...	27 Sicherheit
	224 Connectivity	22 Connectivity	31 Informat...	21 Informat...
	193 Informationen	19 Auslastung...	27 Connectivity	18 Service...
	158 Auslastung_und_Platzangebot	14 DB_App_und_Website	22 Auslastung...	15 Auslastung...
∑ Rest	377 ...	45 ...	46 ...	33 ...
# Aspects	16,200	1,930	2,095	1,547
# non-Null Asp.	12,139	1,380	2,162	1,163

Table 7: Distribution of top-frequent aspects (aspects partly shortened) in Subtask C data

the extracted untokenized document text. Furthermore, the relevance and the document polarity are annotated as well. For relevant documents, the opinion target expressions (OTE) are grouped as `Opinions`. Each `Opinion` contains the token offsets for the OTE, its aspect and the sentiment polarity. Two examples are given in Listing 2. The first one has identifiable OTEs, while the second one – although relevant – does not provide an explicit opinion target expression.

To increase participation and lower the entry boundary, we also provide an TSV format for document-level annotation in order to enable straightforward use with any document classifier. The TSV format contains the following tab-separated fields:

- document id (URL)
- document text
- relevance (true or false)
- document-level polarity, neutral for irrelevant documents
- aspects with polarities; several mentions are possible, empty for irrelevant documents
  Example: *Atmosphäre#Haupt:neutral Atmosphäre#Lautstärke:negative*

Since there are only document-level labels for the TSV format, Subtask D is not evaluated for TSV submissions.

4 Evaluation Measures and Baselines

We evaluate the system predictions using a micro-averaged F1 score. This metric is well-suited for datasets with a clear majority class because each instance is weighted the same as every other one.

For Subtasks A and B (relevance and document-level polarity), we only report the F1 score. For the aspect identification (Subtask C), we report scores for the aspect identification itself, as well as the combination of aspect and sentiment, as it can differ between several aspects in a document. Opinion target expression matching is evaluated in an exact setting, where the token offsets have to match exactly, and a less strict setting, which considers overlaps and partial matches as correct. In detail, we consider an expression a match if the span is +/- one token of the gold data.

Majority Class Baseline The majority class baseline (MCB) yields already a quite good performance, since Subtasks A, B and C have a clear majority class. Thus, the majority class is a strong prior or fallback alternative for instances without much evidence for the other classes. For Subtask C we assign exactly one opinion with the aspect *Allgemein* and the sentiment *neutral*. Since Subtask D is a sequence tagging task, there is no meaningful majority class baseline.

Baseline System We provided a baseline system that uses machine learning with a basic feature set to show the improvements put forth by the participating system. It uses a linear SVM classifier for Subtasks A, B and C and a CRF classifier for the OTEs in Subtask D, both with a minimal set of standard features. The baseline system is available for the participants for initial evaluation and a possible weakly-informed classifier in an ensemble learning setting.[3] Furthermore, it is open-source, so that

[3]The system is available under the Apache Software License 2.0 at: `https://github.com/uhh-lt/`

```
<Document id="http://www.neckar-chronik.de/Home/nachrichten/ueberregional/baden-
    wuerttemberg\_artikel,-Bald-schneller-mit-der-Bahn-von-Deutschland-nach-Paris-\
    _arid,319757.html">
  <text>Bald schneller mit der Bahn von Deutschland nach Paris 5 Stunden 40 Minuten,
      statt wie bisher 6 Stunden 20 Minuten. Stra ß burg. Man kann auch öfter fahren
     . Den neuen grenzüberschreitenden Fahrplan stellte die Regionaldirektion der
     französischen Bahn SNCF am</text>
  <relevance>true</relevance>
  <polarity>positive</polarity>
  <Opinions>
    <Opinion aspect="Zugfahrt#Fahrtzeit_und_Schnelligkeit" from="5" to="14" polarity
       ="positive" target="schneller"/>
    <Opinion category="Zugfahrt#Streckennetz" from="141" to="153" polarity="positive
       " target="öfter fahren"/>
  </Opinions>
</Document>

<Document id="http://twitter.com/majc14055/statuses/649275540877254656">
  <text>@Cmbln Sollte die S- Bahn Berlin nicht einheitlich 80 fahren, wegen
     Konzernvorgabe? Da soll noch Einer durchblicken. ;-)</text>
  <relevance>true</relevance>
  <polarity>neutral</polarity>
  <Opinions>
    <Opinion aspect="Allgemein#Haupt" from="0" to="0" polarity="neutral" target="
       NULL"/>
  </Opinions>
</Document>
```

Listing 2: Example documents in XML format

participants could use parts – like the document readers or the feature extractors – as parts in their systems.

The SVM classifiers use the term frequencies of document terms and a sentiment lexicon (Waltinger, 2010) for prediction. The CRF classifier uses the surface token without processing (lemmatization, standardization, lowercasing) and the POS tag. For tokenization and POS tagging, we use the DKPro tools (Eckart de Castilho and Gurevych, 2014) in the UIMA framework (Ferrucci and Lally, 2004). We have also developed a full system in the course of the same project where the data was annotated, described in (Ruppert et al., 2017). While the full organizer's system did not compete in the shared task as it was developed over a much longer time, it would have been positioned second and third in Subtask A, first and third in Subtask B and first in Subtasks C and D.

5 Participation

Overall, 8 teams participated in the shared task. All of them participated in Subtask B and 5 of them in Subtask A. Only Lee et al. (2017) and Mishra et al. (2017) have participated in Subtasks C and D. Table 8 gives an overview of which team participated in which subtask.

GermEval2017-Baseline

5.1 Participant's Approaches

Across all subtasks, the participants have applied a large variety of approaches. However, we can identify trends and commonalities between the teams, which will be discussed in more detail below. For a detailed description of the approaches, we refer to the referenced papers.

Preprocessing Although some teams have used off-the-shelf tokenizers, such as Schulz et al. (2017) who used the opennlp maxent tokenizer, most of the teams relied on their own implementations. These tokenizers were often combined with large sets of rules that cover social media specific language phenomena such as emoticons, URLs, or repeated punctuation (Sayyed et al., 2017; Sidarenka, 2017; Mishra et al., 2017; Hövelmann and Friedrich, 2017). It would have been possible to use tokenizers from the 2016 EMPIRIST task, e.g. (Remus et al., 2016). Moreover, one team (Hövelmann and Friedrich, 2017) further normalized the data by using an off-the-shelf spell checker and rules to replace e.g. numbers, dates, and URLs with a special token.

Besides a tokenizer, many recent neural classifiers do not require deeper preprocessing. Never-

Team reference	Team name	Subtask			
		A	B	C	D
Schulz et al. (2017)	hda		✓		
UH-HHU-G [4]	UH-HHU-G	✓	✓		
Lee et al. (2017)	UKP_Lab_TUDA	✓	✓	✓	✓
Mishra et al. (2017)	im+sing	✓	✓	✓	✓
Sayyed et al. (2017)	IDS_IULC	✓	✓		
Sidarenka (2017)	PotTS		✓		
Naderalvojoud et al. (2017)	HU-HHU		✓		
Hövelmann and Friedrich (2017)	fhdo	✓	✓		

Table 8: Teams and subtask participation

theless, some of the participants used lemmatizers, chunkers, and part-of-speech taggers (Sidarenka, 2017; Schulz et al., 2017; Naderalvojoud et al., 2017), relying on the TreeTagger by Schmid (1994) or on the Stanford CoreNLP library (Manning et al., 2014).

To compensate for imbalances in the class distribution, two teams have used sampling techniques (Sayyed et al. (2017) and UH-HHU-G) – namely adaptive synthetic sampling (He et al., 2008) and synthetic minority over-sampling (Chawla et al., 2002).

Sentiment Lexicons Most teams used or experimented with some form of word polarity resources. Two teams (Schulz et al., 2017; Sidarenka, 2017) relied on SentiWS (Remus et al., 2010). The resource was also considered but not included in the actual submissions of Hövelmann and Friedrich (2017). Two teams (Schulz et al., 2017; Mishra et al., 2017) have used the lexicon created by Waltinger (2010). Other similarly used resources include the Zurich Polarity List (Clematide and Klenner, 2010) or the LWIC tool (Tausczik and Pennebaker, 2010).

In addition to the use of pre-calculated or manual resources, some teams also created their own lexicons. For instance, Naderalvojoud et al. (2017) created a sense based sentiment lexicon from a large subtitle corpus. Sidarenka (2017) created several lexicons e.g. based on other pre-existing dictionaries and using a German Twitter snapshot.

Dense Word Vectors In addition to word polarity, several teams made use of dense word vectors (also known as word embeddings) and thus integrated distributional semantic word information in their systems. Mishra et al. (2017) trained dense word vectors on large corpus of parliament speeches using GloVe (Pennington et al., 2014). Lee et al. (2017) used word2vec (Mikolov et al., 2013) trained word embeddings on Wikipedia. They also trained sentence vectors on the same data and experimented with German-English bilingual embeddings. Finally, some of the teams relied on fastText (Bojanowski et al., 2017) that makes use of sub-word information to create word vectors, addressing phenomena such as German single-token compounding.

Classifiers When analyzing the classification algorithms utilized by the participants, we identify three major strands. The first strand are approaches that use engineered features to represent the data together with more traditional classification algorithms. The second strand translates the training data in sequences of vectors and feeds them into neural networks. Third, there are ensemble approaches that orchestrate several neural and/or non-neural approaches.

Within the non-neural strand we observe the usage of SVMs (Sidarenka, 2017), CRFs (Mishra et al., 2017; Lee et al., 2017), and threshold based classification (Schulz et al., 2017). Approaches of the neural strand used several different neural network architectures. Most dominant is the usage of recurrent neural networks that contain long-short-term-memory (LSTM) units (UH-HHU-G). In particular, many teams used biLSTM - a variant of LSTMs in which both preceding and following context is considered (Sidarenka, 2017; Mishra et al., 2017; Naderalvojoud et al., 2017; Lee et al., 2017). Other used architectures include convolution layers (UH-HHU-G) and other forms of structured or multi-layered perceptrons (Mishra et al., 2017). Within the ensemble approaches there is an approach of orchestrating several neural networks (Lee et al., 2017), one that combines LSTM and SVM (Sidarenka, 2017), one that uses fastText (Hövelmann and Friedrich, 2017) and two approaches that rely on gradient boosted trees (Hövel-

[4] Submission withdrawn after reviewing

mann and Friedrich, 2017; Sayyed et al., 2017).

6 Evaluation Results

As expected, we observe an increasing difficulty between the subtasks in alphabetical order. This means Subtask A is solved better than B, B solved better than C and C is solved better than D. Interestingly, for all tasks we only see small differences between synchronic and diachronic test sets. From this, we can conclude that either all models are robust against temporary fluctuation, or the distributions in this data do not change at a very high speed. Furthermore, both the majority class baseline and our simple baseline system are quite competitive in all tasks. The detailed results of the subtasks are discussed below.

6.1 Subtask A - Relevance

Most of the teams that participated in Subtask A have beaten the majority class baseline and the baseline system. Table 9 gives an overview of the results. Note that the majority class baseline (0.816) and baseline system (0.852) in this subtask are quite strong. The best system by Sayyed et al. (2017) surpassed our baseline system by 0.05 percent point by using gradient boosted trees and feature selection to obtain the predictions. The second-best team (Hövelmann and Friedrich, 2017) used fastText and applied extensive preprocessing. In future research, it seems worthwhile to examine how these strategies contribute to each system. In addition, we note that the neural approaches of Mishra et al. (2017) and Lee et al. (2017) are almost en par (~ -0.02).

6.2 Subtask B - Document-level Polarity

Similar to Subtask A, we also observe strong baselines in Subtask B, yet that most participants surpass them. Table 10 shows the results. Performance among the top three teams is highly similar. This is particularly interesting as the top three teams Naderalvojoud et al. (2017) [0.749], Hövelmann and Friedrich (2017)[0.748] and Sidarenka (2017) [0.745] have all followed completely different approaches. Naderalvojoud et al. (2017) [0.749] made use of a large lexicon that was combined with a neural network. As already described above, Hövelmann and Friedrich (2017)[0.748] used fastText and extensive preprocessing of the data, whereas Sidarenka (2017) relied on a biLSTM/SVM ensemble. The more or less pure neural approaches of Lee et al. (2017), Sidarenka (2017), and UH-HHU-G yield a slightly worse performance, but still outperform our simple baseline system. Overall, we do not observe large difference on the synchronic versus the diachronic test set, however, most systems marginally lose performance on the diachronic data.

6.3 Subtask C - Aspect-level Polarity

Table 11 shows the performance of the two teams that participated in the aspect-based subtask. Only (Lee et al., 2017) could outperform both provided baselines on the synchronic data. However, the improvements of 0.001 for aspect classification and 0.03 for aspect and sentiment classification are only slight. Surprisingly, on the diachronic data both teams could neither significantly outperform the baseline system nor the majority class baseline (*Allgemein:neutral*). Interestingly, we observe a increased performance for all submitted runs for the diachronic data.

6.4 Subtask D - Target Extraction

The same teams that worked on Subtask C also participated in Subtask D. Both teams relied on neural network approaches and outperformed both baselines. While the structured perceptron of Mishra et al. (2017) achieved the best results for the exact metric, the combination of LSTM and CRF by Lee et al. (2017) gained the – by far – best results for the overlap metric. In Table 12 we report the results. As expected, the results of the overlap metric are better than those of the exact metric, as the exact metric is more strict. Similar to subtask C, we can conclude that the diachronic data can be classified more easily in both metrics.

7 Related Work

First of all, our shared task is related to shared tasks on aspect-based sentiment analysis that were conducted within the international workshop on semantic evaluation (SemEval) (Pontiki et al., 2014; Pontiki et al., 2015; Pontiki et al., 2016). However, we here focus exclusively on German but target a larger, monolingual data set. We also relate to previous German shared tasks on aspect-based sentiment analysis such as Ruppenhofer et al. (2016). In contrast to this work, we are pursuing an annotation scheme that is inspired by the needs of a industrial customer as opposed to linguistic considerations.

Team	Run	synchronic	diachronic
Sayyed et al. (2017)	xgboost	.903	.906
Hövelmann and Friedrich (2017)	fasttext	.899	.897
Mishra et al. (2017)	biLSTM strucutured perceptron	.879	.870
Lee et al. (2017)	stacked learner CCA SIF embedding	.873	.881
Hövelmann and Friedrich (2017)	gbt_bow	.863	.856
organizers	*baseline system*	*.852*	*.868*
UH-HHU-G	ridge classifier char fourgram	.835	.849
UH-HHU-G	linear SVC l2 char fivegram	.834	.859
UH-HHU-G	passive-aggressive char fivegram	.827	.850
UH-HHU-G	linear SVC l2 trigram	.824	.837
organizers	*majority class baseline*	*.816*	*.839*
UH-HHU-G	gru mt	.816	.840
UH-HHU-G	cnn gru sent mt	.810	.839
Hövelmann and Friedrich (2017)	ensemble	.734	.160

Table 9: Results for Subtask A on relevance detection.

Team	Run	synchronic	diachronic
Naderalvojoud et al. (2017)	SWN2-RNN	.749	.736
Hövelmann and Friedrich (2017)	fasttext	.748	.742
Sidarenka (2017)	bilstm-svm	.745	.718
Naderalvojoud et al. (2017)	SWN1-RNN	.737	.736
Sayyed et al. (2017)	xgboost	.733	.750
Sidarenka (2017)	bilstm	.727	.704
Lee et al. (2017)	stacked learner CCA SIF embedding	.722	.724
Hövelmann and Friedrich (2017)	gbt_bow	.714	.714
Hövelmann and Friedrich (2017)	ensemble	.710	.725
UH-HHU-G	ridge classifier char fourgram	.692	.691
Mishra et al. (2017)	biLSTM strucutured perceptron	.685	.675
UH-HHU-G	linear SVC l2 char fivegram	.680	.692
organizers	*baseline system*	*.667*	*.694*
UH-HHU-G	linearSVC l2 trigram	.663	.702
organizers	*majority class baseline*	*.656*	*.672*
UH-HHU-G	gru mt	.656	.672
UH-HHU-G	cnn gru sent mt	.644	.668
Schulz et al. (2017)		.612	.616
UH-HHU-G	Passive-Aggressive char fivegram	.575	.676

Table 10: Results for Subtask B on sentiment detection.

As we are examining directed opinions, we also relate to shared tasks that were conducted on automatically detecting stance from social media data. Stance is defined as being in favor or against a given target, which can be a politician, a political assertion or any controversial issue. Stance detection has been addressed by a couple of recent shared tasks – namely SemEval 2016 task 6 (Mohammad et al., 2016), NLPCC Task 4 (Xu et al., 2016) or IBEREVAL 2017 (Taulé et al., 2017). Similar to them, we find that state-of-the-art methods still have a long way to go to solve the problem and that, in contrast to other domains and tasks, neural networks are not clearly superior and often inferior to more traditional rule-based or feature engineering approaches.

8 Conclusions

In this paper, we describe a shared task on aspect-based sentiment analysis in social media customer feedback. Our shared task includes four subtasks, in which the participants had to detect A) whether feedback is relevant to the given topic *Deutsche Bahn*, B) which overall sentiment is express by a review, C) what aspects are evaluated, and D) what linguistic expressions are used to express these aspects. We provide an annotated data set of almost 28,000 messages from several social media sources. Thereby our dataset represents the largest set of German sentiment annotated reviews.

The shared task attracted a high variance of approaches from 8 different teams. We observe that the usage of gradient boosted trees, large sentiment lexicons, and the connection of neural and more traditional classifiers are advantageous strategies

Team	Run	synchronic		diachronic	
		aspect	aspect + sentiment	aspect	aspect + sentiment
Lee et al. (2017)	LSTM CRF stacked learner correct offsets	.482	.354	-	-
organizers	*baseline system*	.481	.322	.495	.389
organizers	*majority class baseline*	.442	.315	.456	.384
Mishra et al. (2017)	biLSTM strucutured perceptron	.421	.349	.460	.401
Lee et al. (2017)	LSTM CRF stacked learner correct offsets 2	.358	.308	-	-
Lee et al. (2017)	LSTM-CRF only correct offsets	.095	.081	-	-

Table 11: Results for Subtask C on aspect-based sentiment detection.

Team	Run	synchronic		diachronic	
		exact	overlap	exact	overlap
Mishra et al. (2017)	biLSTM strucutured perceptron	.220	.221	.281	.282
Lee et al. (2017)	LSTM CRF stacked learner correct offsets	.203	.348	-	-
Lee et al. (2017)	LSTM CRF stacked learner correct offsets 2	.186	.267	-	-
organizers	*baseline system*	*.170*	*.237*	*.216*	*.271*
Lee et al. (2017)	LSTM-CRF only correct offsets	.089	.089	-	-
Lee et al. (2017)	LSTM-CRF stacked learner 4 polarity correct offsets	.024	.183	-	-

Table 12: Results for Subtask D on opinion target expression identification

for the formulated tasks. Nevertheless, our simple baseline classifier is highly competitive across all tasks. We will release the annotated dataset as part of this task. This will hopefully strengthen the research on German sentiment and social media analysis.

Acknowledgments

We would like to thank Axel Schulz and Maria Pelevina from the Deutsche Bahn Fernverkehr AG for their collaboration in the project *ABSA-DB: Aspect-based Sentiment Analysis for DB Products and Services* as part of the Innovation Alliance DB & TU Darmstadt. We would also like to thank Ji-Ung Lee for his helpful feedback. In addition, this work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media". Finally, we thank the Interest Group on German Sentiment Analysis (IGGSA) for their endorsement.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Simon Clematide and Manfred Klenner. 2010. Evaluation and extension of a polarity lexicon for german. In *Proceedings of the First Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 7–13, Lisbon, Portugal.

Richard Eckart de Castilho, Eva Mujdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In *Proceedings of the LT4DH workshop at COLING 2016*, pages 76–84, Osaka, Japan.

Richard Eckart de Castilho and Iryna Gurevych. 2014. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proc. Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT) at COLING 2014*, pages 1–11, Dublin, Ireland.

David Ferrucci and Adam Lally. 2004. UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Natural Language Engineering 2004*, 10(3-4):327–348.

Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pages 1322–1328, Hong Kong, China. IEEE.

Leonard Hövelmann and Christoph M. Friedrich. 2017. Fasttext and Gradient Boosted Trees at GermEval-2017 Tasks on Relevance Classification and Document-level Polarity. In *Proceedings of the*

GSCL GermEval Shared Task on Aspect-based Sentiment in Social Media Customer Feedback, pages 30–35, Berlin, Germany.

Ji-Ung Lee, Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. UKP TU-DA at GermEval 2017: Deep Learning for Aspect Based Sentiment Detection. In *Proceedings of the GSCL GermEval Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, pages 22–29, Berlin, Germany.

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, Baltimore, MD, USA.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Workshop at International Conference on Learning Representations (ICLR).*, pages 1310–1318, Scottsdale, AZ, USA.

Pruthwik Mishra, Vandan Mujadia, and Soujanya Lanka. 2017. GermEval 2017 : Sequence based Models for Customer Feedback Analysis. In *Proceedings of the GSCL GermEval Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, pages 36–42, Berlin, Germany.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the International Workshop on Semantic Evaluation 2016*, San Diego, USA.

Behzad Naderalvojoud, Behrang Qasemizadeh, and Laura Kallmeyer. 2017. HU-HHU at GermEval-2017 Sub-task B: Lexicon-Based Deep Learning for Contextual Sentiment Analysis. In *Proceedings of the GSCL GermEval Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, pages 18–21, Berlin, Germany.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, Austin, TX, USA.

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 SemEval 2014*, pages 27–35, Dublin, Ireland.

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 SemEval*, pages 486–495, Denver, Colorado.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphee De Clercq, Veronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Núria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th SemEval*, pages 19–30, San Diego, California.

Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. SentiWS-A Publicly Available German-language Resource for Sentiment Analysis. In *Proceedings of the 7th International Language Ressources and Evaluation (LREC'10)*, pages 1168–1171, Valletta, Malta.

Steffen Remus, Gerold Hintz, Chris Biemann, Christian M. Meyer, Darina Benikova, Judith Eckle-Kohler, Margot Mieskes, and Thomas Arnold. 2016. EmpiriST: AIPHES-Robust Tokenization and POS-Tagging for Different Genres. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X)*, pages 106–114, Berlin, Germany.

Josef Ruppenhofer, Julia Maria Struß, and Michael Wiegand. 2016. Overview of the IGGSA 2016 Shared Task on Source and Target Extraction from Political Speeches. In *Bochumer Linguistische Arbeitsberichte*, pages 1–9, Bochum, Germany.

Eugen Ruppert, Abhishek Kumar, and Chris Biemann. 2017. LT-ABSA: An extensible open-source system for document-level and aspect-based sentiment analysis. In *Proceedings of the GSCL GermEval Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, pages 55–60, Berlin, Germany.

Zeeshan Ali Sayyed, Daniel Dakota, and Sandra Kübler. 2017. IDS-IUCL Contribution to GermEval 2017. In *Proceedings of the GSCL GermEval Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, pages 43–48, Berlin, Germany.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Karen Schulz, Margot Mieskes, and Christoph Becker. 2017. h-da Participation at GermEval Subtask B: Document-level Polarity. In *Proceedings of the GSCL GermEval Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, pages 13–17, Berlin, Germany.

Uladzimir Sidarenka. 2017. PotTS at GermEval-2017 Task B: Document-Level Polarity Detection Using Hand-Crafted SVM and Deep Bidirectional LSTM Network. In *Proceedings of the GSCL GermEval Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, pages 49–54, Berlin, Germany.

Mariona Taulé, M Antonia Martı, Francisco Rangel, Paolo Rosso, Cristina Bosco, and Viviana Patti. 2017. Overview of the task of Stance and Gender Detection in Tweets on Catalan Independence at IBEREVAL 2017. In *Notebook Papers of 2nd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL)*, volume 19, Murcia, Spain.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Ulli Waltinger. 2010. Sentiment Analysis Reloaded: A Comparative Study On Sentiment Polarity Identification Combining Machine Learning And Subjectivity Features. In *Proceedings of the 6th International Conference on Web Information Systems and Technologies (WEBIST '10)*, pages 203–210, Valencia, Spain.

Ruifeng Xu, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. 2016. Overview of NLPCC Shared Task 4: Stance Detection in Chinese Microblogs. In *International Conference on Computer Processing of Oriental Languages*, pages 907–916, Kunming, China. Springer.